**PROJECT REPORT**

(Project Term Jul-Dec 2022)

**Diabetes Detection Using Machine Learning**

## 

Submitted by

**(Ravi Pandey ) Registration Number :11910459**

**Course Code INT248**

Under the Guidance of

##### 

# 

Niharika Thakur

# **School of Computer Science and Engineering**

Text

Description automatically generated with low confidence

**DECLARATION**

We hereby declare that the project work entitled (“ **Diabetes Detection Using Machine Learning** ”) is an authentic record of our own work carried out as requirements of Project for the award of B.Tech degree in Computer science from Lovely Professional University, Phagwara, under the guidance of (Niharika Thakur), during Jul to Dec 2022. All the information furnished in this project report is based on our own intensive work and is genuine.

Name of Student : Ravi Pandey

Registration Number: 11910459

(Ravi Pandey)

Date: 03-11-2022

**CERTIFICATE**

This is to certify that the declaration statement made by this group of students is correct to the best of my knowledge and belief. They have completed this Project under my guidance and supervision. The present work is the result of their original investigation, effort and study. No part of the work has ever been submitted for any other degree at any University. The Project is fit for the submission and partial fulfillment of the conditions for the award of B.Tech degree in Computer Science from Lovely Professional University, Phagwara.

**Signature and Name of the Mentor**

**Designation**

**School of Computer Science and Engineering,**

Lovely Professional University,

Phagwara, Punjab.

Date :03-11-2022

**ACKNOWLEDGEMENT**

I would sincerely like to thanks for the constructive criticism, support, encouragement

valuable, comments, suggestions, timely helps and many innovative ideas given to

me by my project supervisor Niharika Thakur in carrying out project and the report

I must convey my gratitude to Niharika Thakur for giving me the constant source of

Inspiration and help in preparing the project ,personally correcting my work and

Providing encouragement throughout the project.

I also thanks all my faculty members for steering me through the tough as well as

Easy phase of the project in result oriented manner with concern attention.

ABSTRACT

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis.

According to International Diabetes Federation 382 million people are living with diabetes

across the whole world. By 2035, this will be doubled as 592 million. Diabetes mellitus or

imply diabetes is a disease caused due to the increase level of blood glucose. Various

traditional methods, based on physical and chemical tests, are available for diagnosing

diabetes. However, early prediction of diabetes is quite challenging task for medical

practitioners due to complex interdependence on various factors as diabetes affects human

organs such as kidney, eye, heart, nerves, foot etc. Data science methods have the potential to

benefit other scientific fields by shedding new light on common questions. One such task is

to help make predictions on medical data. Machine learning is an emerging scientific field in

data science dealing with the ways in which machines learn from experience. The aim of this

project is to develop a system which can perform early prediction of diabetes for a patient

with a higher accuracy by combining the results of different machine learning techniques.

This project aims to predict diabetes via three different supervised machine learning methods

including: SVM, Logistic regression, KNN. This project also aims to propose an effective

technique for earlier detection of the diabetes disease using Machine learning algorithms and

end to end deployment using flask.

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis.According to International Diabetes Federation 382 million people are living with diabetesacross the whole world. By 2035, this will be doubled as 592 million. Diabetes mellitus orimply diabetes is a disease caused due to the increase level of blood glucose. Varioustraditional methods, based on physical and chemical tests, are available for diagnosingdiabetes. However, early prediction of diabetes is quite challenging task for medicalpractitioners due to complex interdependence on various factors as diabetes affects humanorgans such as kidney, eye, heart, nerves, foot etc. Data science methods have the potential tobenefit other scientific fields by shedding new light on common questions. One such task isto help make predictions on medical data. Machine learning is an emerging scientific field indata science dealing with the ways in which machines learn from experience. The aim of thisproject is to develop a system which can perform early prediction of diabetes for a patientwith a higher accuracy by combining the results of different machine learning techniques.This project aims to predict diabetes via three different supervised machine learning methodsincluding: SVM, Logistic regression, KNN. This project also aims to propose an effectivetechnique for earlier detection of the diabetes disease using Machine learning algorithms andend to end deployment using flask

CONTENTS

1. INTRODUCTION

2. RELATED WORKS

3. DATASET

4. PROPOSED METHODS

I) DATASET COLLECTION

II) DATA PRE-PROCESSING

III) MISSING VALUE IDENTIFICATION

IV) FEATURE SELECTION:

V) SCALING AND NORMALIZATION

VI) SPLITTING OF DATA

VII) DESIGN AND IMPLEMENTATION OF CLASSIFICATION MODEL

VIII) MACHINE LEARNING CLASSIFIER

5. MODELING AND ANALYSIS

a) LOGISTIC REGRESSION

b) K-NEAREST NEIGHBORS:

c) SVM

d) NAIVE BAYES:

e) DECISION TREE:

f) RANDOM FOREST

g) ADABOOST CLASSIFIER

6. MEASURMENT

7. RESULTS AND DISCUSSION

8. RESULTS AND ANALYSIS

9. CONCLUSION

10. REFERENCES

1. Introduction

All around there are numerous ceaseless infections that are boundless in evolved anddeveloping nations. One of such sickness is diabetes. Diabetes is a metabolic issue thatcauses blood sugar by creating a significant measure of insulin in the human body or byproducing a little measure of insulin. Diabetes is perhaps the deadliest sickness on the planet.It is not just a malady yet, also a maker of different sorts of sicknesses like a coronaryfailure, visual deficiency, kidney ailments and nerve harm, and so on.Subsequently, the identification of such chronic metabolic ailment at a beginningperiod could help specialists around the globe in forestalling loss of human life. Presently,with the ascent of machine learning, AI, and neural systems, and their application in variousdomains [1, 2] we may have the option to find an answer for this issue. ML strategies andneural systems help scientists to find new realities from existing well-being-relatedinformational indexes, which may help in ailment supervision and detection. The currentwork is completed utilizing the Pima Indians Diabetes Database. The point of this frameworkis to make an ML model, which can anticipate with precision the likelihood or the odds of apatient being diabetic. The ordinary distinguishing process for the location of diabetes is thatthe patient needs to visit a symptomatic focus. One of the key issues of bio-informaticsexamination is to achieve precise outcomes from the information. Human mistakes orvarious laboratory tests can entangle the procedure of identification of the disease. Thismodel can foresee whether the patient has diabetes or not, aiding specialists to ensure thatthe patient in need of clinical consideration can get it on schedule and also help anticipate theloss of human lives.DNA makes neural networks the apparent choice. Neural networks use neurons to transmitdata across various layers, with each node working on a different weighted parameter to helppredict diabetes. Presently, with the ascent of machine learning, AI, and neural systems, and theirapplication in various domains [1, 2] we may have the option to find an answer for this issue

1. Related work

Diabetes prediction is a classification technique with two mutually exclusivepossibleoutcomes, either the person is diabetic or not diabetic. After extensive research,wecame to conclusion that although numerous classification techniques can be used forthepurpose of prediction, the observed accuracy varied. On careful examination oftheperformance of techniques used in prevalent works, logistic regression, KNN,Naive Bayes[3], random forest, decision tree, and neural network [4], we found themat par when appliedto our dataset. KNN and logistic regression techniques were ableto achieve 80% accuracy.The primary factor which influenced our algorithm selection was its adaptabilityandcompatibility with future applications. The inevitable shift of data storage towardDNAmakes neural networks the apparent choice. Neural networks use neurons totransmit dataacross various layers, with each node working on a different weightedparameter to helppredict diabetes.The point of this framework is to make an ML model, which can anticipatewith precision thelikelihood or the odds of a patient being diabetic. The ordinarydistinguishing process for thelocation of diabetes is that the patient needs to visit asymptomatic focus. One of the keyissues of bio-informatics examination is to achieveprecise outcomes from the information.Human mistakes or various laboratory testscan entangle the procedure of identification of thedisease. This model can foreseewhether the patient has diabetes or not, aiding specialists toensure that the patientin need of clinical consideration can get it on schedule and also helpanticipate theloss of human lives

1. Data Set

The dataset collected is originally from the Pima Indians Diabetes Database isavailable on Kaggle. It consists of several medical analyst variables and one target variable.The objective of the dataset is to predict whether the patient has diabetes or not. The datasetconsists of several independent variables and one dependent variable, i.e., the outcome.Independent variables include the number of pregnancies the patient has had their BMI,insulin level, age, and so on as Shown in Following Table 1:

Table 1 Dataset description

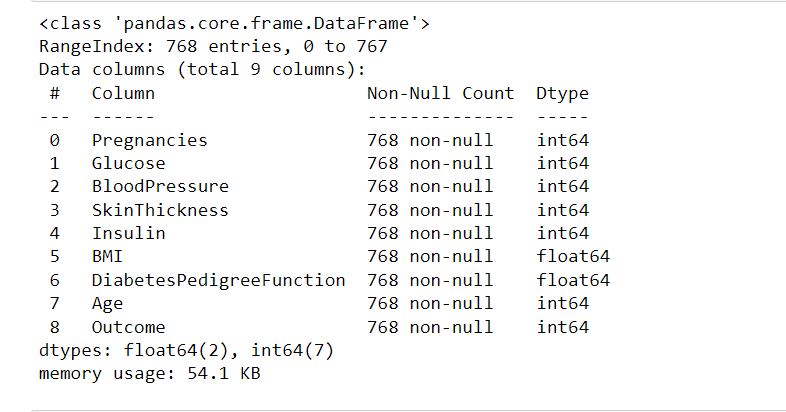
|  |  |  |
| --- | --- | --- |
| Serial no | Attribute Names | Description |
| 1 | Pregnancies | Number of 琀椀mes pregnant |
| 2 | Glucose | Plasma glucose concentra琀椀on |
| 3 | Blood Pressure | Diastolic blood pressure |
| 4 | Insulin | 2-h serum insulin |
| 5 | BMI | Body mass index |
| 6 | Diabetes pedigree function | Diabetes pedigree func琀椀on |
| 7 | Outcome | Class variable (0 or 1) |
| 8 | Age | Age of pa琀椀ent |
| 9 | Skin Thickness | Triceps skin fold thickness (mm)  5  Triceps skin fold thickness (mm)  5  Triceps skin fold thickness (mm)  5  Triceps skin fold thickness |

Fig 3.1 dataset description.

➔ The diabetes data set consists of 2000 data points, with 9 features each.

➔ “Outcome” is the feature we are going to predict, 0 means No diabetes, 1 means diabetes.

Fig 3.2



predictions➔ There is no null values in dataset.

Correlation Matrix:

Chart, treemap chart

Description automatically generated

n Mrix:

Fig 3.3 correla琀椀on matrix

2. Related Works

Diabetes prediction is a classification technique with two mutually exclusive

possibleoutcomes, either the person is diabetic or not diabetic. After extensive research,

wecame to conclusion that although numerous classification techniques can be used forthe

purpose of prediction, the observed accuracy varied. On careful examination ofthe

performance of techniques used in prevalent works, logistic regression, KNN,Naive Bayes

[3], random forest, decision tree, and neural network [4], we found themat par when applied

to our dataset. KNN and logistic regression techniques were ableto achieve 80% accuracy.

The primary factor which influenced our algorithm selection was its adaptabilityand

compatibility with future applications. The inevitable shift of data storage towardDNA

makes neural networks the apparent choice. Neural networks use neurons totransmit data

across various layers, with each node working on a different weightedparameter to help

predict diabetes.

The point of this framework is to make an ML model, which can anticipatewith precision the

likelihood or the odds of a patient being diabetic. The ordinarydistinguishing process for the

location of diabetes is that the patient needs to visit asymptomatic focus. One of the key

issues of bio-informatics examination is to achieveprecise outcomes from the information.

Human mistakes or various laboratory testscan entangle the procedure of identification of the

disease. This model can foreseewhether the patient has diabetes or not, aiding specialists to

ensure that the patientin need of clinical consideration can get it on schedule and also help

anticipate theloss of human lives

2. Related Works

Diabetes prediction is a classification technique with two mutually exclusive

possibleoutcomes, either the person is diabetic or not diabetic. After extensive research,

wecame to conclusion that although numerous classification techniques can be used forthe

purpose of prediction, the observed accuracy varied. On careful examination ofthe

performance of techniques used in prevalent works, logistic regression, KNN,Naive Bayes

[3], random forest, decision tree, and neural network [4], we found themat par when applied

to our dataset. KNN and logistic regression techniques were ableto achieve 80% accuracy.

The primary factor which influenced our algorithm selection was its adaptabilityand

compatibility with future applications. The inevitable shift of data storage towardDNA

makes neural networks the apparent choice. Neural networks use neurons totransmit data

across various layers, with each node working on a different weightedparameter to help

predict diabetes.

The point of this framework is to make an ML model, which can anticipatewith precision the

likelihood or the odds of a patient being diabetic. The ordinarydistinguishing process for the

location of diabetes is that the patient needs to visit asymptomatic focus. One of the key

issues of bio-informatics examination is to achieveprecise outcomes from the information.

Human mistakes or various laboratory testscan entangle the procedure of identification of the

disease. This model can foreseewhether the patient has diabetes or not, aiding specialists to

ensure that the patientin need of clinical consideration can get it on schedule and also help

anticipate theloss of human lives 2. Related WorksDiabetes prediction is a classification technique with two mutually exclusivepossibleoutcomes, either the person is diabetic or not diabetic. After extensive research,wecame to conclusion that although numerous classification techniques can be used forthepurpose of prediction, the observed accuracy varied. On careful examination oftheperformance of techniques used in prevalent works, logistic regression, KNN,Naive Bayes[3], random forest, decision tree, and neural network [4], we found themat par when appliedto our dataset. KNN and logistic regression techniques were ableto achieve 80% accuracy.The primary factor which influenced our algorithm selection was its adaptabilityandcompatibility with future applications. The inevitable shift of data storage towardDNAmakes neural networks the apparent choice. Neural networks use neurons totransmit dataacross various layers, with each node working on a different weightedparameter to helppredict diabetes.The point of this framework is to make an ML model, which can anticipatewith precision thelikelihood or the odds of a patient being diabetic. The ordinarydistinguishing process for thelocation of diabetes is that the patient needs to visit asymptomatic focus. One of the keyissues of bio-informatics examination is to achieveprecise outcomes from the information.Human mistakes or various laboratory testscan entangle the procedure of identification of thedisease. This model can foreseewhether the patient has diabetes or not, aiding specialists toensure that the patientin need of clinical consideration can get it on schedule and also helpanticipate theloss of human lives

* 1. correlation matrix

It is easy to see that there is no single feature that has a very high correlation with ouroutcome value. Some of the features have a negative correlation with the outcome value andsome have positive

Correlation Matrix:

Fig 3.3 correla琀椀on matrix

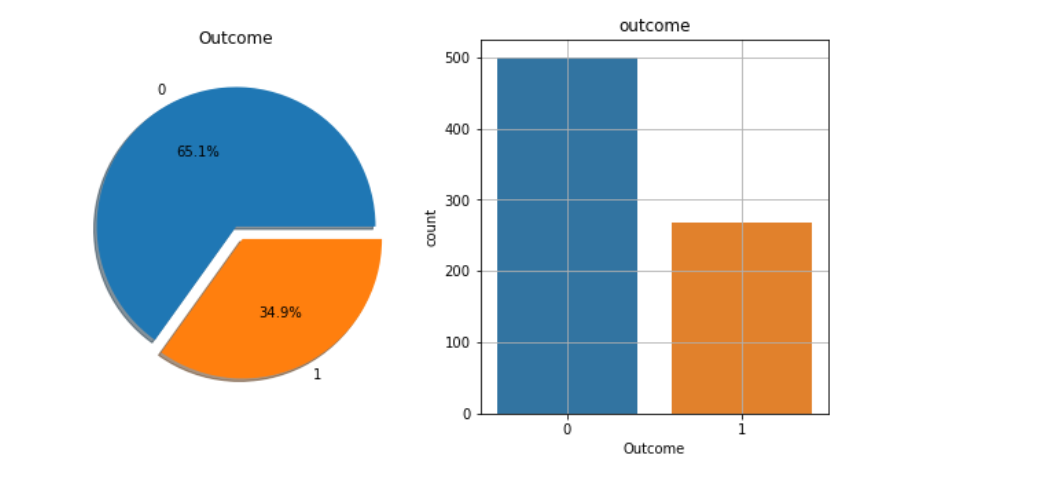
It is easy to see that there is no single feature that has a very high correlation with our

outcome value. Some of the features have a negative correlation with the outcome value and

some have positive.

Skew Of Data:

Pia & Bar Plot for Outcome Class



The above graph shows that the data is biased towards datapoints having outcome value as 0

where it means that diabetes was not present actually. The number of non-diabetics is almost

twice the number of diabetic patients

The above graph shows that the data is biased towards datapoints having outcome value as 0 where it means that diabetes was not present actually. The number of non-diabetics is almost twice the number of diabetic patients.

1. PROPOSED METHODS

I] Dataset collection – It includes data collection and understanding the data to study the

hidden patterns and trends which helps to predict and evaluating the results. Dataset carries

1405 rows i.e., total number of data and 10 columns i.e., total number of features. Features

include Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI,

DiabetesPedigreeFunction, Ag

I] Dataset collection – It includes data collection and understanding the data to study thehidden patterns and trends which helps to predict and evaluating the results. Dataset carries1405 rows i.e., total number of data and 10 columns i.e., total number of features. Featuresinclude Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI,DiabetesPedigreeFunction, Age.

II] Data Pre-processing:

This phase of model handles inconsistent data in order to get moreaccurate and precise results like in this dataset Id is inconsistent so we dropped the feature.This dataset doesn’t contain missing values. So, we imputed missing values for few selectedattributes like Glucose level, Blood Pressure, Skin Thickness, BMI and Age because theseattributes cannot have values zero. Then data was scaled using StandardScaler. Since therewere a smaller number of features and important for prediction so no feature selection wasdone.III]Missing value identification: Using the Panda library and SK-learn , we got the missing values in the datasets, shown in Table 2. We replaced the missing value with the corresponding mean value.



IV] Feature selection:

Pearson’s correlation method is a popular method to find the most

vrelevant attributes/features. The correlation coefficient is calculated in this method, which

correlates with the output and input attributes. The coefficient value remains in the range by

between −1 and 1. The value above 0.5 and below −0.5 indicates a notable correlation, and

the zero value means no correlation

IV] Feature selection:

Pearson’s correlation method is a popular method to find the mostvrelevant attributes/features. The correlation coefficient is calculated in this method, whichcorrelates with the output and input attributes. The coefficient value remains in the range bybetween −1 and 1. The value above 0.5 and below −0.5 indicates a notable correlation, andthe zero value means no correlation

V] Scaling and Normalization: We performed feature scaling by normalizing the data from 0 to 1 range, which boosted thealgorithm’s calculation speed.scaling means that you're transforming your data so that it fits within a specific scale, like 0-100 or 0-1. You want to scale data when you're using methods based on measures of how farapart data points are, like support vector machines (SVM) or k-nearest neighbours (KNN).With these algorithms, a change of "1" in any numeric feature is given the same importance.

VI] Spliting of data:

After data cleaning and pre-processing, the dataset becomes ready to

train and test. In the train/split method, we split the dataset randomly into the training and

testing set. For Training we took 1600 sample and for testing we took 400 sample.

After data cleaning and pre-processing, the dataset becomes ready totrain and test. In the train/split method, we split the dataset randomly into the training andtesting set. For Training we took 1600 sample and for testing we took 400 sample.

VII] Design and implementation of classification model: In this research work, comprehensive studies are done by applying different ML classification techniques like DT, KNN, RF, NB, LR, SVM.

VIII] Machine learning classifier:We have developed a model using Machine learning Technique. Used different classifier andensemble techniques to predict diabetes dataset. We have applied SVM, LR, DT and RFMachine learning classifier to analyse the performance by finding accuracy of each classifierAll the classifiers are implemented using scikit learn libraries in python. The implementedclassification algorithms are described in next section.

5 MODELING AND ANALYSIS

A] Logistic Regression:Logistic regression is a machine learning technique used when dependent variables are able to categorize. The outputs obtained by using the logistic regression is based on the available features. Here sigmoidal function is used to categorize the output.

B] K-Nearest Neighbors:K-nearest neighbors (KNN) algorithm uses ‘feature similarity’ to predict the values of newdatapoints which further means that the new data point will be assigned a value based onhow closely it matches the points in the training set. Predictions are made for a new instance (x) by searching through the entire training set forthe K most similar instances (the neighbors) and summarizing the output variable for those Kinstances.

C]SVM:SVM is supervised learning algorithm used for classification. In SVM we have to identifythe right hyper plane to classify the data correctly. In this we have to set correct parametersvalues. To find the right hyper plane we have to find right margin for this we have choose thegamma value as 0.0001 and rbf kernel. If we select the hyper plane with low margin leads tomiss classification.

D] Naive Bayes: Naive Bayes classifiers are a collection of classification algorithms based on Bayes’Theorem. It is not a single algorithm but a family of algorithms where all of them share acommon principle, i.e. every pair of features being classified is independent of each other.

E] Decision Tree:Decision tree is non parametric classifier in supervised learning. In this method all the detailsare represented in the form of tree, where leaves are corresponds to the class labels andattributes are corresponds to internal node of the tree. We have used Gini Index for splittingthe nodes.

F] Random Forest: Random forest is an ensemble learning method for classification. This algorithm consists oftrees and the number of tree structures present in the data is used to predict the accuracy.Where leaves are corresponds to the class labels and attributes are corresponds to internalnode of the tree. Here number of trees in forest used is 100 in number and Gini index is usedfor splitting the nodes.

1. Measurements

To find the effienct classifier for diabetes prediction we have applied a performance matricesare confusion matrix and accuracy are discussed as follows:Confusion matrix: - which provides output matrix with complete description performance of the model.Here, Tp: True positive FP: False positive TN: True negative FN: False negative.

1. RESULTS AND DISCUSSION

Machine learning classification algorithms developed for prediction of diabetes in earlierstage. We used 70% of data for trining and 30% of data for testing. In this ratio of datasplitting Here we found that Random Forest Classifier predicted with 99% of accuracy ashighest accuracy for the dataset. Comparison of results of all the implemented classifiers arelisted in below. ResultsCreating a User Interface for Accessibility:

The last part of the project is the creation of a user interface for the model. This userinterface is used to enter unseen data for the model to read and then make a prediction. Theuser interface is created using “Flask” Web app, Hyper Text Markup Language, andCascading Style Sheets.Machine Learning Algorithms Result Logistic Regression 79.0 ,K-Nearest Neighbors 80.5 ,SVM 84.5 ,Naive Bayes 76.83 ,Decision Tree 96.0 Random Forest 98.0, AdaBoost Classifier 81.1

1. Results and Analysis

The project predicts the onset of diabetes in a person based on the relevant medical detailscollected. When the person enters all the relevant medical data required in the online Webportal, this data is then passed on to the trained model for it to make predictions whether theperson is diabetic or non-diabetic the model then makes the prediction with an accuracy of98%, which is fairly good and reliable. Following figure shows the basic UI form whichrequires the user to enter the specific medical data fields. These parameters help determine ifthe person is prone to develop diabetes Our research has the added benefit of an associatedWeb app, which makes the model more user friendly and easily understandable for a novice

1. Conclusion

The objective of the project was to develop a model which could identify patients withdiabetes who are at high risk of hospital admission. Prediction of risk of hospital admissionis a fairly complex task. Many factors influence this process and the outcome. There ispresently a serious need for methods that can increase healthcare institution’s understandingof what is important in predicting the hospital admission risk. This project is a smallcontribution to the present existing methods of diabetes detection by proposing a system thatcan be used as an assistive tool in identifying the patients at greater risk of being diabetic.This project achieves this by analyzing many key factors like the patient’s blood glucoselevel, body mass index, etc., using various machine learning models and throughretrospective analysis of patients’ medical records. The project predicts the onset of diabetesin a person based on the relevant medical details that are collected using aWebapplication.When the user enters all the relevant medical data required in the online Webapplication, this data is then passed on to the trained model for it to make predictionswhether the person is diabetic or nondiabetic. The model is developed using artificial neuralnetwork consists of total of six dense layers. Each of these layers is responsible for theefficient working of the model. The model makes the prediction with an accuracy of 98%,which is fairly good and reliable.

5. MODELING AND ANALYSIS:

A] Logistic Regression:

Logistic regression is a machine learning technique used when dependent variables are able

to categorize. The outputs obtained by using the logistic regression is based on the available

features. Here sigmoidal function is used to categorize the output.

B] K-Nearest Neighbors:

K-nearest neighbors (KNN) algorithm uses ‘feature similarity’ to predict the values of new

datapoints which further means that the new data point will be assigned a value based on

how closely it matches the points in the training set.

Predictions are made for a new instance (x) by searching through the entire training set for

the K most similar instances (the neighbors) and summarizing the output variable for those K

instances.

C]SVM:

SVM is supervised learning algorithm used for classification. In SVM we have to identify

the right hyper plane to classify the data correctly. In this we have to set correct parameters

values. To find the right hyper plane we have to find right margin for this we have choose the

gamma value as 0.0001 and rbf kernel. If we select the hyper plane with low margin leads to

miss classification.

D] Naive Bayes:

Naive Bayes classifiers are a collection of classification algorithms based on Bayes’

Theorem. It is not a single algorithm but a family of algorithms where all of them share a

common principle, i.e. every pair of features being classified is independent of each other.

E] Decision Tree:

Decision tree is non parametric classifier in supervised learning. In this method all the details

are represented in the form of tree, where leaves are corresponds to the class labels and

attributes are corresponds to internal node of the tree. We have used Gini Index for splitting

the nodes.

F] Random Forest:

Random forest is an ensemble learning method for classification. This algorithm consists of

trees and the number of tree structures present in the data is used to predict the accuracy.

Where leaves are corresponds to the class labels and attributes are corresponds to internal

node of the tree. Here number of trees in forest used is 100 in number and Gini index is use

2. Related Works

Diabetes prediction is a classification technique with two mutually exclusive

possibleoutcomes, either the person is diabetic or not diabetic. After extensive research,

wecame to conclusion that although numerous classification techniques can be used forthe

purpose of prediction, the observed accuracy varied. On careful examination ofthe

performance of techniques used in prevalent works, logistic regression, KNN,Naive Bayes

[3], random forest, decision tree, and neural network [4], we found themat par when applied

to our dataset. KNN and logistic regression techniques were ableto achieve 80% accuracy.

The primary factor which influenced our algorithm selection was its adaptabilityand

compatibility with future applications. The inevitable shift of data storage towardDNA

makes neural networks the apparent choice. Neural networks use neurons totransmit data

across various layers, with each node working on a different weightedparameter to help

predict diabetes.

The point of this framework is to make an ML model, which can anticipatewith precision the

likelihood or the odds of a patient being diabetic. The ordinarydistinguishing process for the

location of diabetes is that the patient needs to visit asymptomatic focus. One of the key

issues of bio-informatics examination is to achieveprecise outcomes from the information.

Human mistakes or various laboratory testscan entangle the procedure of identification of the

disease. This model can foreseewhether the patient has diabetes or not, aiding specialists to

ensure that the patientin need of clinical consideration can get it on schedule and also help

anticipate theloss of human lives  
?

ABSTRACT

Diabetes Prediction:

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis.

According to International Diabetes Federation 382 million people are living with diabetes

across the whole world. By 2035, this will be doubled as 592 million. Diabetes mellitus or

imply diabetes is a disease caused due to the increase level of blood glucose. Various

traditional methods, based on physical and chemical tests, are available for diagnosing

diabetes. However, early prediction of diabetes is quite challenging task for medical

practitioners due to complex interdependence on various factors as diabetes affects human

organs such as kidney, eye, heart, nerves, foot etc. Data science methods have the potential to

benefit other scientific fields by shedding new light on common questions. One such task is

to help make predictions on medical data. Machine learning is an emerging scientific field in

data science dealing with the ways in which machines learn from experience. The aim of this

project is to develop a system which can perform early prediction of diabetes for a patient

with a higher accuracy by combining the results of different machine learning techniques.

This project aims to predict diabetes via three different supervised machine learning methods

including: SVM, Logistic regression, KNN. This project also aims to propose an effective

technique for earlier detection of the diabetes disease using Machine learning algorithms and

end to end deployment using flask

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis.

According to International Diabetes Federation 382 million people are living with diabetes

across the whole world. By 2035, this will be doubled as 592 million. Diabetes mellitus or

imply diabetes is a disease caused due to the increase level of blood glucose. Various

traditional methods, based on physical and chemical tests, are available for diagnosing

diabetes. However, early prediction of diabetes is quite challenging task for medical

practitioners due to complex interdependence on various factors as diabetes affects human

organs such as kidney, eye, heart, nerves, foot etc. Data science methods have the potential to

benefit other scientific fields by shedding new light on common questions. One such task is

to help make predictions on medical data. Machine learning is an emerging scientific field in

data science dealing with the ways in which machines learn from experience. The aim of this

project is to develop a system which can perform early prediction of diabetes for a patient

with a higher accuracy by combining the results of different machine learning techniques.

This project aims to predict diabetes via three different supervised machine learning methods

including: SVM, Logistic regression, KNN. This project also aims to propose an effective

technique for earlier detection of the diabetes disease using Machine learning algorithms and

end to end deployment using flask.

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis.

According to International Diabetes Federation 382 million people are living with diabetes

across the whole world. By 2035, this will be doubled as 592 million. Diabetes mellitus or

imply diabetes is a disease caused due to the increase level of blood glucose. Various

traditional methods, based on physical and chemical tests, are available for diagnosing

diabetes. However, early prediction of diabetes is quite challenging task for medical

practitioners due to complex interdependence on various factors as diabetes affects human

organs such as kidney, eye, heart, nerves, foot etc. Data science methods have the potential to

benefit other scientific fields by shedding new light on common questions. One such task is

to help make predictions on medical data. Machine learning is an emerging scientific field in

data science dealing with the ways in which machines learn from experience. The aim of this

project is to develop a system which can perform early prediction of diabetes for a patient

with a higher accuracy by combining the results of different machine learning techniques.

This project aims to predict diabetes via three different supervised machine learning methods

including: SVM, Logistic regression, KNN. This project also aims to propose an effective

technique for earlier detection of the diabetes disease using Machine learning algorithms and

end to end deployment using flask