# DIABETES PREDICTION

By Kandula Bharath Reddy

# ABSTRACT

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to the 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 simply diabetes is a disease caused due to the increased 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 a 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, ANN. This project also aims to propose an effective technique for earlier detection of the diabetes disease. Keywords: Machine Learning, Supervised, Svm, Ann, Logistic Regression.

# INTRODUCTION

Diabetes is one of the deadliest diseases in the world. It is not only a disease but also a creator of different kinds of diseases like heart attack, blindness, kidney diseases, etc. 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. Moreover, every time they want to get their diagnosis report, they have to waste their money in vain.Diabetes Mellitus (DM) is defined as a group of metabolic disorders mainly caused by abnormal insulin secretion and/or action. Insulin deficiency results in elevated blood glucose levels (hyperglycemia) and impaired metabolism of carbohydrates, fat and proteins. DM is one of the most common endocrine disorders, affecting more than 200 million people worldwide. The onset of diabetes is estimated to rise dramatically in the upcoming years. DM can be divided into several distinct types. However, there are two major clinical types, type 1 diabetes (T1D) and type 2 diabetes (T2D), according to the etiopathology of the disorder. T2D appears to be the most common form of diabetes (90% of all diabetic patients), mainly characterized by insulin resistance. The main causes of T2D include lifestyle, physical activity, dietary habits and heredity, whereas T1D is thought to be due to autoimmune logical destruction of the Langerhans islets hosting pancreatic-β cells. T1D affects almost 10% of all diabetic patients worldwide, with 10% of them ultimately developing idiopathic diabetes. Other forms of DM, classified on the basis of insulin secretion profile and/or onset, include Gestational Diabetes, endocrinopathies, MODY (Maturity Onset Diabetes of the Young), neonatal, mitochondrial, and pregnancy diabetes. The symptoms of DM include polyuria, polydipsia, and significant weight loss among others. Diagnosis depends on blood glucose levels (fasting plasma glucose = 7.0 mmol/L.

Machine learning is the scientific field dealing with the ways in which machines learn from experience. For many scientists, the term “machine learning” is identical to the term “artificial intelligence”, given that the possibility of learning is the main characteristic of an entity called intelligent in the broadest sense of the word. The purpose of machine learning is the construction of computer systems that can adapt and learn from their experience. A more detailed and formal definition of machine learning is given by Mitchel: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. 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. 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 Artificial Neural Network algorithms.

## Relevance of the Project:

In recent years, using Machine Learning has been used with increasing frequency to predict the possibility of disease. Many algorithms and toolkits have been created and studied by researchers. These have highlighted the tremendous potential of this research field. Based on several studies, we found that a commonly used dataset was the Pima Indians Diabetes Dataset from the University of California, Irvine (UCI) Machine Learning Database. In this project logistic regression aimed at validating a chosen class label of given data and aimed at building the final classifier model. All the studies presented above used the same Pima Indians Diabetes Dataset as the experimental material.

## Purpose:

Large amount of data has been continuously generated in the field of engineering and science. Recent advances in technology have resulted in big electronic data that allow data to be captured, processed, analyzed and stored rather inexpensively. This change leads to new trends in market as well as industry such as Internet banking and e-commerce, insurance, financial transactions, supermarket, healthcare, communications, location of data that generate huge amounts of electronic data. The need to understand huge, complex, information rich data sets is important to virtually all fields in business, science, engineering and medical. The data used in data warehouses and data marts has been extracted from knowledge hidden in that data. This knowledge is becoming vital in today's increasingly competitive world. The greatest problem of today is how to teach people to ignore irrelevant data. With the rise of Machine Learning approaches we have the ability to find a solution to diabetes prediction, we have developed a system using machine learning which has the ability to predict whether the patient has diabetes or not. Furthermore, predicting the disease early leads to treating the patients before it becomes critical. Machine Learning 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 logistic regression.

## Scope of the Project:

The early intervention of diabetes can reduce the prevalence of diabetes and hence the economic burden due to it. Machine Learning techniques play an important role in treatment plan workout, rehabilitation, chronic diseases management plan etc. Long term follow up plans may be easily guided and keen supervision is possible. The systems may definitely be helpful in reducing the cost of patient management by avoiding unnecessary investigations and patient follow up. These prediction systems will add accuracy and time management. Computer-based patient support systems benefit patients by providing informational support that increases their participation in health care.

## Problem Statement and Definition:

To identify whether a given person in the dataset will be diabetic, non diabetic or pre- diabetic will be done on the basis of attribute values. Values exceeding a specific value may contribute to identifying whether a person is diabetic, non diabetic or pre-diabetic. The aim of prediction of diabetes is to make people aware about diabetes and what it takes to treat it and gives the power to control. The model can be used by the endocrinologists, dietitians, ophthalmologists and podiatrists to predict if or if not the patient is likely to suffer from diabetes, if yes, how intense it could be. The dataset consists of features comprising the medical details of the patients that are useful in determining the health condition of the patient.

## Goals of the System:

- Convert manual to computerised. Before this, the majority of the process is done manually. After converting it to computerize it will be easy to predict.
- By computerizing, it is easier to understand the doctor's report which is hardly understood with different and complicated handwriting.
- Easy to maintain record.
- Enable to predict various types of diabetes.
- Ensure the system is useful to doctors and patients.

## Objective of the study:

The primary objective of this assignment is to broaden a platform so as to be simple and smooth to apply, as right here one has to provide the patient’s scientific details and primarily based on the features extracted the algorithm will then discover

diabetes and its type. As right here set of rules does the task hence a well trained version is much less certain to make errors in predicting diabetes and its type consequently, in short accuracy is advanced and thereby it additionally saves time and makes simpler for doctors in addition to sufferers to expect whether or not they may be vulnerable to any type of diabetes or not, that is otherwise we difficult to do without health practitioner's involvement. No human intervention required: To predict diabetes one must provide scientific details which includes age, BMI, and so on. And right here the set of rules will offer the effects based on the capabilities extracted and consequently here probabilities of mistakes being made are very minimal given that there is no human intervention and it also saves a lot of time for the sufferers or doctors and they could similarly continue for treatments or different tactics should quicker. This is in case whilst consequences are provided quicker to them. This can in-turn make the precaution/prevention for diabetes faster while it saves medical doctors and affected persons the essential time, to be able to cross on to in addition treatments and precautions to be taken to minimize the effect of that diabetes. Not the simplest hit upon the diabetes kind but additionally suggest precautions: In this mission our goal isn't only to find and are expecting the kind of diabetes but pin point towards the precautions to be taken to minimize the impact of the diabetes. Getting hints on precautions to be taken will help the doctors and sufferers to progress without problems to similar steps of their treatment.

## Existing system:

The healthcare enterprise collects big quantities of healthcare facts which, unluckily, are not "mined" to find out hidden information. Clinical decisions are frequently made based totally on docs‟ intuition and enjoyment instead of on the knowledge of wealthy statistics hidden in the database. This exercise ends in unwanted biases, errors and excessive medicine. The existing process is very slow to give the result. It is very difficult to find diabetes or not.

## Proposed system:

Diabetes prediction is an internet-primarily based device gaining knowledge of utility, skilled through a Pima Indian dataset. The person inputs its particular clinical information to get the prediction of diabetes. The set of rules will calculate the opportunity of presence of diabetes. Thus, minimizing the price and time required to be expecting the disorder. Format of statistics plays an essential element in this software. At the time of uploading the user information utility will take a look at its right record format and if it is no longer as consistent with want then ERROR dialog box may be induced. Our device might be implementing the algorithm: Logistic Regression. The algorithms may be educated using the statistics set obtained from University of California, Irvine. 75% of the entries in the statistics set can be used for

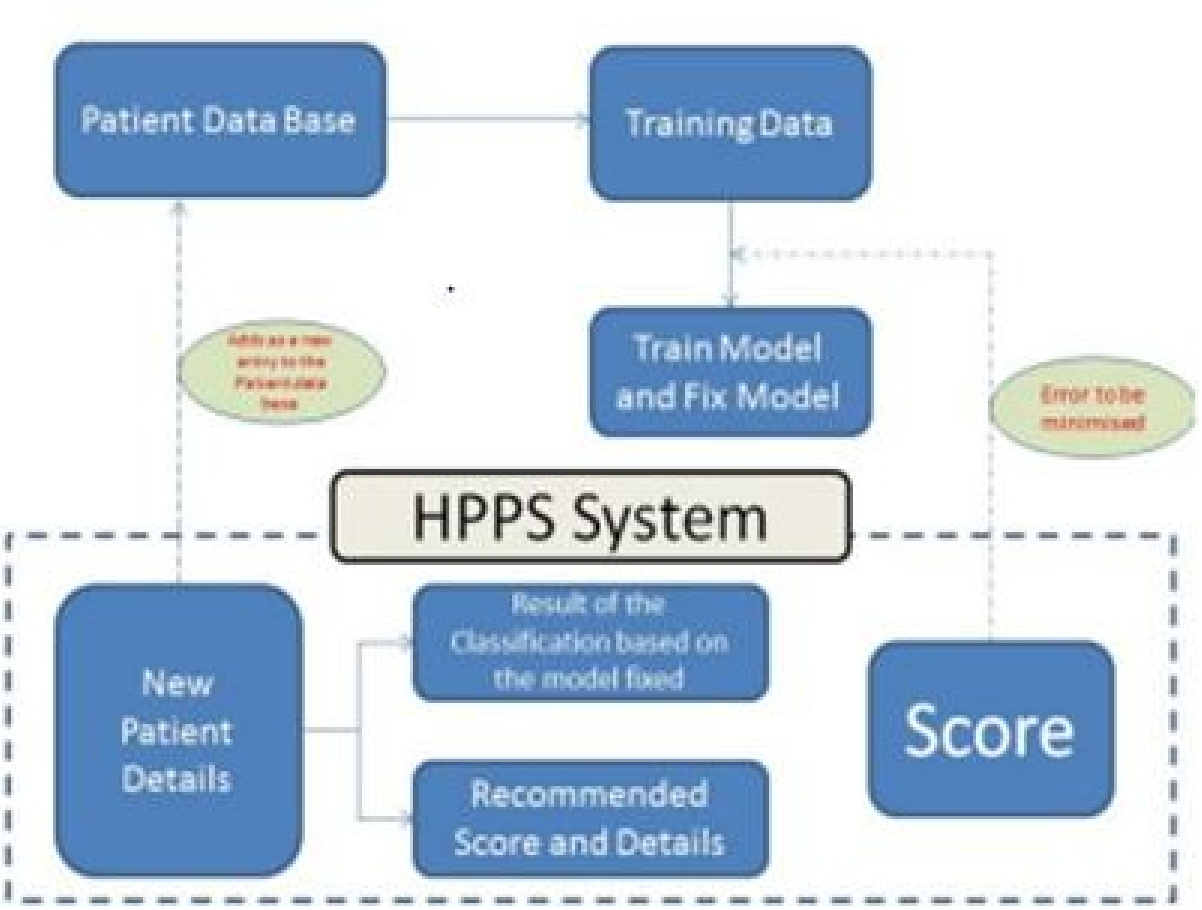
education and the last 25% for testing the accuracy of the set of rules. Furthermore, a few steps can be taken for optimizing the algorithms thereby enhancing the accuracy.

## Advantages of the system:

Our proposed System has the following benefits:

- Powerful, flexible and easy to use.
- Increased efficiency of doctors.
- Improved patient satisfaction.
- Reduce the use of papers.
- Simple and Quick.
- More accurate result.

## Flow-chart:



# LITERATURE SURVEY

The Pima Indians are genetically predisposed to diabetes, and it was noted that their diabetic rate was 19 times that of a typical town in Minnesota. The National Institute of Diabetes and Digestive and Kidney Disease of the National Institute of Health (NIH) originally owned the Pima Indian diabetes Database (PIDD) [18]. The number of patients (n) in the database n=768 each with 9 attribute variables. Out of the nine conditional attributes, six are due to physical examination, the rest of the attributes are chemical examination. Of these 9 attributes, there are eight inputs and the last one being the output. The goal is to use the first 8 variables to predict attribute values of the $9^{th}$ variables.

The aim is to develop a hybrid model for classifying Pima Indian Diabetic Database. The datasets were identified and the incorrectly classified instances were eliminated using K-means clustering. Then using the decision Tree classifier the classification is done on these correctly clustered instances. The resultant dataset is used to train and test the diabetic data set using two methods dividing training data and test data using 60- q40 ratio and 10 fold cross Validation method. Experimental results show the improvement in accuracy of diabetic data set using proposed cascaded method: k-means with decision tree by an order of 19.50% of classification compared to decision tree alone with unprocessed data.

Weka is a popular machine learning software developed in Java at University of Waikato, New Zealand. It is open source software available at GNU (General Public License). It consists of visualization tools and algorithms which are used in data analysis and predictive modeling with graphical user interface for easy functionality access. Weka supports several data mining tasks such as data pre-processing, clustering, regression, visualization and feature selection. The attributes available in Weka are of one of these types: Nominal: one of predefined list of values, Numeric: A real or integer number, String, Date, Relational. Key features of this tool are open source and platform independent. This consists of various algorithms for data mining and machine learning.

Diabetes, a non-communicable disease, is leading to long-term complications and serious health problems. A report from the World Health Organisation addresses diabetes and its complications that impact on individuals physically, financially, economically over the families. The survey says about 1.2 million deaths due to the uncontrolled stage of health lead to death. About 2.2 million deaths occurred due to the risk factors of diabetes like cardiovascular and other diseases.

Diabetes is an ailment caused due to the extended level of sugar obsession in the blood. In this paper, discussing various classifiers, a decision support system is proposed that uses the AdaBoost algorithm with Decision Stump as a base classifier for classification. Moreover, Support Vector Machine, Naive Bayes and Decision Tree have additionally executed as a base classifiers for AdaBoost calculation for exactness confirmation. The accuracy obtained for AdaBoost calculation with

choices stump as a base classifier is 80.72%, which is more noteworthy contrasted with that of Support Vector Machine, Naive Bayes and Decision Tree.

Artificial intelligence is having more effect than machine realizing, which creates calculations ready to take in examples and choose standards from information. Machine learning calculations have been implanted into information mining pipelines, which can consolidate them with established measurable techniques, to remove learning from information. Inside the EU-financed MOSAIC undertaking, an information mining pipeline has been utilized to determine an arrangement of prescient models of sort 2 diabetes mellitus (T2DM) entanglements in light of electronic wellbeing record information of almost one thousand patients. Such pipeline includes clinical focus profiling, prescient model focusing on, prescient model development and model approval. In the wake of having managed to miss information by methods for irregular woods (RF) and having connected appropriate methodologies to deal with class unevenness, we have utilized Logistic Regression with the stepwise component choice to foresee the beginning of retinopathy, neuropathy, or nephropathy, at various time situations, at 3, 5, and 7 years from the main visit at the Hospital Center for Diabetes (not from the conclusion). Considered factors are sexual orientation, age, time of determination, weight file (BMI), glycated haemoglobin (HbA1c), hypertension, and smoking propensity. Lust models, custom fitted as per the complexities, gave an exact up to 0.838. Diverse factors were chosen for every complexity and time situation, prompting particular models simply to mean the clinical practice.

In this paper, analysis of a Pima Indian dataset is done using various classification techniques like Naïve Bayes, Zero R, J48, random forest, MLP, logistic regression. Comparison and prediction of positive and negative diabetes. Diagnosing diabetes through a data mining tool using the WEKA tool, in terms of accuracy and performance MLP is better.

Patients with diabetes should ceaselessly screen their blood glucose levels and modify insulin measurements, endeavouring to keep blood glucose levels as near typical as would be prudent. Blood glucose levels that veer off from the typical range can prompt genuine here and now and long-haul intricacies. A programmed expectation shows that cautioned individuals of fast approaching changes in their blood glucose levels would empower them to make a preventive move. In this paper, we depict an answer that uses a bland physiological model of blood glucose progression to produce enlightening highlights for a support vector regression display that is prepared with tolerant particular information. The new model beats diabetes specialists at foreseeing blood glucose levels and could be utilized to envision right around a fourth of hypoglycaemic occasions 30 min ahead of time. In spite of the fact that the comparing exactness is right now only 42%, most false cautions are in close hypoglycaemic locales and hence patients reacting to these hypoglycaemia alarms would not be hurt by intercession.

Diabetes mellitus is a standout amongst the most genuine wellbeing challenges in both creating and created nations. As per the International Diabetes Federation, there are 285 million diabetic individuals around the world. This

aggregate is relied upon to ascend to 380 million in 20 years. Because of its significance, an outline of a classifier for the recognition of Diabetes ailment with ideal cost and better execution is the need of the age. The Pima Indian diabetic database at the UCI machine learning research facility has turned into a standard for testing information mining calculations to see their expectation exactness in diabetes information arrangement. The proposed strategy utilizes SVM, a machine learning technique as the classifier for analysis of diabetes. The machine learning strategy centres around arranging diabetes illness from a high dimensional therapeutic dataset. The trial came about to demonstrate that a help vector machine can be effectively utilized for diagnosing diabetes illness.

The point of this examination is to the finding of diabetes illness, which is a standout amongst the most vital infections in the restorative field utilizing Generalized Discriminant Analysis (GDA) and Least Square Support Vector Machine (LS-SVM). Likewise, we proposed another course learning framework in light of Generalized Discriminant Analysis and Least Square Support Vector Machine. The proposed framework comprises two phases. The primary stage, we have utilized Generalized Discriminant Analysis to discriminate highlight factors amongst sound and patient (diabetes) information as a pre-preparing process. The second stage, we have utilized LS-SVM so as to order the diabetes dataset. While LS-SVM acquired 78.21% grouping precision utilizing 10-overlap. cross approval, the proposed framework called GDA–LS-SVM got 82.05% order exactness utilizing 10-crease across approval. The heartiness of the proposed framework is inspected utilizing arrangement precision, k-crease cross-approval technique and disarray lattice. The acquired order exactness is 82.05% and it is exceptionally encouraging contrasted with the beforehand detailed grouping strategies.

# TECHNICAL REQUIREMENTS OF THE SYSTEM

## Artificial Neural Network:

The artificial neural network is much similar to the natural neural network of a brain. Artificial Neural networks (ANN) typically consist of multiple layers or a cube design, and the signal path traverses from front to back. Back propagation is the use of forward stimulation to reset weights on the "front" neural units and this is sometimes done in combination with training where the correct result is known. More modern networks are a bit freer flowing in terms of stimulation and inhibition with connections interacting in a much more chaotic and complex fashion. Dynamic neural networks are the most advanced, in that they dynamically can, based on rules, for new connections and even new neural units while disabling others.Generally, the artificial neural network is consisting of the layers and network function, the layers of the network are including: input layer, hidden layer and output layer. The input neurons define all the input attribute values for the data mining model. In our work, the number of neurons is 7, since each item in our data set has 7 attributes, including: Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and age. For the hidden layer, hidden neurons receive inputs from input neurons and provide outputs to output neurons. The hidden layer is where the various probabilities of the inputs are assigned weights. A weight describes the relevance or importance of a particular input to the hidden neuron. Mathematically, a neuron's network function f(x) is defined as composition of other functions gi (x), which can further be defined as a composition of other functions. The important characteristic of the activation function is that it provides a smooth transition as input values change, like a small change in input produces a small change in output. The artificial neural networks are applied to tend to fall within the broad categories. Application areas include the system identification and control (vehicle control, trajectory prediction, process control, natural resources management), quantum chemistry, game playing and decision making (backgammon, chess, poker), pattern recognition (radar systems, face identification, object recognition and more), sequence recognition (gesture, speech, handwritten text recognition), medical diagnosis, financial applications (e.g. automated trading systems), data mining (or knowledge discovery in databases, "KDD"), visualization and e-mail spam filtering.

Artificial neural networks have also been used to diagnose several cancers. An ANN based hybrid lung cancer detection system named HLND improves the accuracy of diagnosis and the speed of lung cancer radiology. These networks have also been used to diagnose prostate cancer. The diagnoses can be used to make specific models taken from a large group of patients compared to information of one

given patient. The models do not depend on assumptions about correlations of different variables. Colorectal cancer has also been predicted using the neural networks. Neural networks could predict the outcome for a patient with colorectal cancer with more accuracy than the current clinical methods. After training, the networks could predict multiple patient outcomes from unrelated institutions.

## Support Vector Machine:

The Support Vector Machine (SVM) was first proposed by Vapnik, and SVM is a set of related supervised learning methods always used in medical diagnosis for classification and regression. SVM simultaneously minimises the empirical classification error and maximises the geometric margin. So SVM is called Maximum Margin Classifiers. SVM is a general algorithm based on guaranteed risk bounds of statistical learning theory, so called structural risk minimization principle. SVMs can efficiently perform nonlinear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. The kernel trick allows constructing the classifier without explicitly knowing the feature space.for binary classification tasks, so we choose SVM to predict the diabetes. The reason is SVM is well known for its discriminative power for classification, especially in the cases where a large number of features are involved, and in our case where the dimension of the feature is 7.

## Logistic Regression:

In statistics Logistic regression is a regression model where the dependent variable is categorical, namely binary dependent variable-that is, where it can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick. Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences. For example, the Trauma and Injury Severity Score (TRISS), which is widely used to predict mortality in injured patients, was originally developed using logistic regression. Many other medical scales used to assess severity of a patient have been developed using logistic regression. The technique can also be used in engineering, especially for predicting the probability of failure of a given process, system or product. It is also used in marketing applications such as prediction of a customer's propensity to purchase a product or halt a subscription. In economics it can be used to predict the likelihood of a person's choosing to be in the labor force, and a business application is about to predict the likelihood of a homeowner defaulting on a mortgage. Conditional random fields, an extension of logistic regression to sequential data, are used in natural language processing. In this paper, Logistic regression was used to predict whether a patient suffered from diabetes, based on seven observed characteristics of the patient.

# IMPLEMENTATION

## Main Application:

```
import numpy as np
from flask import Flask, request, render_template
import pickle

app = Flask(__name__)
sc = pickle.load(open('sc.pkl', 'rb'))
model = pickle.load(open('classifier.pkl', 'rb'))


@app.route('/')
def home():
    return render_template('ditect.html')

@app.route('/predict',methods=['POST'])
def predict():

    float_features = [float(x) for x in request.form.values()]
    final_features = [np.array(float_features)]
    pred = model.predict( sc.transform(final_features) )
    return render_template('result.html', prediction = pred)

if __name__ == "__main__":
    app.run(debug=True)
```

## Data Analysis code:

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns

"""### Data Collection"""


#Extracting data

dataset=pd.read_csv('diabetes.csv')
dataset.head()
```

```
#Our dataset dimensions

dataset.shape
dataset.describe()

#Counting values of outcomes having 0 or 1, 0 means non diabetic and 1 means diabetic

sns.countplot(x='Outcome',data=dataset)
dataset['Outcome'].value_counts()
dataset.groupby('Outcome').mean()

#Correlation matrix to show correlation between two variables, 0.x means x% similar
corr_mat=dataset.corr()

sns.heatmap(corr_mat, annot=True)

#Ex: correlation between Glucose and Outcome is 47% that means output depends majorly on Glucose

""" ### Data Cleaning"""


#Check if any null or empty data is present in dataset

dataset.isna().sum()

#Feature matrix - Taking all our independent columns into single array and dependent values into another array

x=dataset.iloc[:,:-1].values  #Independent matrix
y=dataset.iloc[:,-1].values
x.shape
x[0] #refering to column 1 in dataset i.e pregnancies
y

"""### Exploratory Data Analysis #### Checking which columns are useful or not"""


#glucose for diabetic

fig = plt.figure(figsize =(16,6))
sns.distplot(dataset["Glucose"][dataset["Outcome"] == 1])
```

```
plt.xticks([i for i in range(0,201,15)],rotation = 45)
plt.ylabel("Glucose count")
plt.title("Glucose",fontsize = 20)

#insulin for diabetic

fig = plt.figure(figsize = (16,6))
sns.distplot(dataset["Insulin"][dataset["Outcome"]==1])
plt.xticks()
plt.title("Insulin",fontsize = 20)

#BMI for diabetic

fig = plt.figure(figsize =(16,6))
sns.distplot(dataset["BMI"][dataset["Outcome"]==1])
plt.xticks()
plt.title("BMI",fontsize = 20)

#diabetes pedigree function for diabetic

fig = plt.figure(figsize = (16,5))
sns.distplot(dataset["DiabetesPedigreeFunction"][dataset["Outcome"] == 1])
plt.xticks([i*0.15 for i in range(1,12)])
plt.title("diabetes pedigree function")

#Age for diabetic

fig = plt.figure(figsize = (16,6))
sns.distplot(dataset["Age"][dataset["Outcome"] == 1])
plt.xticks([i*0.15 for i in range(1,12)])
plt.title("Age")

#Removing unnecessary columns

x = dataset.drop(["Pregnancies","BloodPressure","SkinThickness","Outcome"],axis =
1)
y = dataset.iloc[:,-1]

#splitting dataset into training set and test set

from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=0)

#test_size 0.2 means for testing data 20% and training data 80%
```

```
x_train.shape #80% of original dataset (769,9) after removing unnecceasry data
x_test.shape #20% of original dataset (769,9) after removing unnecceasry data

#Feature Scaling - To standardize the independent features present in the data in a fixed range.
#If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller
#values as the lower values, regardless of the unit of the values.

from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
x_train = sc.fit_transform(x_train)
x_test = sc.transform(x_test)
x_train


"""### Model Building - K Nearest Neighbor"""

from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors =25, metric = 'minkowski')

#n_neighbors is 25 bcoz for x_train we got 614 which is near to 25^2
#metric means on what factor choosing so as its KNN so our metric is minkowski i.e., distance

knn.fit(x_train, y_train)

#Predicting the data

knn_y_pred = knn.predict(x_test)
knn_y_pred

# Confusion matrix - To check how many are correct or wrong

from sklearn.metrics import confusion_matrix
knn_cm = confusion_matrix(y_test, knn_y_pred)
sns.heatmap(knn_cm, annot=True)

# The above heatmap says 0,0 means true negative and 1,1 means true positive
# and 0,1 means even person is negative but showing result positive
# and 1,0 means person is positive but shows negative so its danger so we need to accurate our model
```

```
print("Correct:",sum(knn_y_pred==y_test))
print("Incorrect : ",sum(knn_y_pred != y_test))
print("Accuracy:",sum(knn_y_pred ==y_test)/len(knn_y_pred))


#Verifying accuracy using inbuilt methods

from sklearn.metrics import accuracy_score
accuracy_score(y_test,knn_y_pred)


"""### Simple Vector Machine"""

from sklearn.svm import SVC
svc=SVC(kernel="linear",random_state=0)
svc.fit(x_train,y_train)
svc_y_pred = svc.predict(x_test)
svc_cm = confusion_matrix(y_test,svc_y_pred)
print(svc_cm)
print("Correct:",sum(svc_y_pred == y_test))
print("Incorrect : ",sum(svc_y_pred != y_test))
print("Accuracy:",sum(svc_y_pred ==y_test)/len(knn_y_pred))


"""### Naive Bias"""

from sklearn.naive_bayes import GaussianNB
nb_classifier = GaussianNB()
nb_classifier.fit(x_train,y_train)
nb_y_pred =nb_classifier.predict(x_test)
nb_cm = confusion_matrix(nb_y_pred,y_test)
print(nb_cm)
print("Correct:",sum(nb_y_pred == y_test))
print("Incorrect : ",sum(nb_y_pred != y_test))
print("Accuracy:",sum(nb_y_pred ==y_test)/len(nb_y_pred))

### Saving the classifier

import pickle
pickle.dump(svc, open('classifier.pkl', 'wb'))
pickle.dump(sc, open('sc.pkl', 'wb'))
```

## Web-page Code:

### Home:

```
<!DOCTYPE html>
<html>
   <head>
   <title>DITECT</title>
   <meta name="viewport" content="width=device-width, initial-scale=1">
   <link rel="stylesheet"
href="https://cdnjs.cloudflare.com/ajax/libs/font-awesome/4.7.0/css/font-awesome.mi
n.css">

   <style>
   body {
   font-family: Arial, Helvetica, sans-serif;
   }
   input[type=text], select {
   width: 100%;
   padding: 12px 20px;
   margin: 8px 0;
   display: inline-block;
   border: 1px solid #ccc;
   border-radius: 4px;
   box-sizing: border-box;
   }
   .registerbtn {
   background-color: #e48912;
   color: white;
   padding: 16px 20px;
   margin: 8px 0;
   border: none;
   cursor: pointer;
   width: 100%;
   opacity: 0.9;
   }
   .registerbtn:hover {
   opacity:1;
   }
   p{
      color: rgb(58, 000, 00);
      font-family:'Lucida Sans', 'Lucida Sans Regular', 'Lucida Grande', 'Lucida Sans
Unicode', Geneva, Verdana, sans-serif;
```

```
    }
    input[type=submit]:hover {
    background-color: #17cce4;
    }
    .center {
    margin: auto;
    width: 60%;
    border: 3px solid #ff9101c5;
    padding: 10px;
    }

    </style>
    </head>
    <body>

    <h2> <p style="text-align:center;">Diabetes Prediction</p></h2>
    <form action="/predict"method="post">
   <!-- <input type="text" name="Pregnancies" placeholder="Pregnancies"
required="required" /> -->
    <input type="text" name="Glucose Level" placeholder="Glucose Level"
required="required" />
    <!--  <input type="text" name="Blood Pressure" placeholder="Blood Pressure"
required="required" />
    <input type="text" name="Skin Thickness" placeholder="Skin Thickness"
required="required" />-->
    <input type="text" name="Insulin" placeholder="Insulin" required="required" />
    <input type="text" name="BMI" placeholder="BMI" required="required" />
    <input type="text" name="Diabetes PF" placeholder="Diabetes PF"
required="required" />
    <input type="text" name="Age" placeholder="Age" required="required" />
    <button type="submit" class="registerbtn">Predict</button>

    </form>
    </div>
    </body>
</html>
```

## Result:

```
<!DOCTYPE html>
<html>
<head>
```

```
  <title>DITECT</title>
<meta name="viewport" content="width=device-width, initial-scale=1">
<link rel="stylesheet"
href="https://cdnjs.cloudflare.com/ajax/libs/font-awesome/4.7.0/css/font-awesome.mi
n.css">

<style>
body {
  font-family: Arial, Helvetica, sans-serif;
}

input[type=text], select {
  width: 100%;
  padding: 12px 20px;
  margin: 8px 0;
  display: inline-block;
  border: 1px solid #ccc;
  border-radius: 4px;
  box-sizing: border-box;
}
.registerbtn {
  background-color: #e48912;
  color: white;
  padding: 16px 20px;
  margin: 8px 0;
  border: none;
  cursor: pointer;
  width: 100%;
  opacity: 0.9;
}
.registerbtn:hover {
  opacity:1;
}
input[type=submit]:hover {
  background-color: #e48912;
}

</style>
</head>
<body>

 <h2>Results: </h2>
{% if prediction == 1%}
```

```
<h2 style="color: rgb(189, 16, 16);">Chances of having Diabetes is more, please consult a Doctor.</h2>
{% elif prediction == 0%}
<h2 style="color: rgb(25, 197, 25);">No Worries!!! You don't have Diabetes.</h2>
{% endif %}

  </body>
  </html>
```

# CONCLUSION and FUTURE SCOPE

## Conclusion:

Diabetes is a vital health hassle in human society. This paper has summarised the kingdom of art techniques and to be had techniques for prediction of this sickness. Deep studying and rising regions of Machine Learning showed a few promising bring about different areas of clinical diagnosis with excessive accuracy. It continues to be an open area waiting to get applied in Diabetes prediction. Some strategies of deep studying have been discussed which may be implemented for Diabetes prediction, alongside pioneer machine getting to know algorithms. An analytical assessment has been completed for locating our best available algorithm for clinical dataset. In future our purpose is to carry ahead the work of temporal scientific dataset, wherein the dataset varies with time and retraining of the dataset is needed.

## Future Scope:

The proposed system is Diabetes prediction. We can enhance this system to predict diseases and suggest medications using machine learning. Further the system can be extended to N number diseases existing with proper medications.