Department of Electronics and
Electrical communications Engineering
Faculty of Engineering
Cairo University

Diabetes Prediction

Submitted to
DR Samah Tantawy

Diabetes prediction Using Random Forest method

Ahmed Mahmoud Abusaif, Ahmed Mahmoud Ali, Ahmed Mokhtar Mahfouz, Hassan Mohamed Hassan, Abdelrahman Mohamed, Abdallah Hossam-Eldin

Abstract-Diabetes is referred to as hyperglycemia or high blood sugar. It is a leading cause of death worldwide and is classified as the fifth leading cause of death in United States. In engineering, we aim to find several solutions to predict diabetes and the other diseases. One of the techniques we found very helpful to solve such problems is the machine learning. We choose random forest because of its high accuracy and it's simple to understand. In random forest we use some features selecting algorithms such as principal component analysis (PCA) and minimum redundancy maximum relevance (MRMR). By using python code we create an application for this problem and its prediction. The accuracy of KNN and decision tree is very low compared to the SVM and random forest.

Key words-Diabetes, Random forest, SVM, Decision Tree and classifications.

I. Introduction

Diabetes is a disease that threatens human life expectancy. A person is suffering from diabetes, when blood sugar levels are above normal. Pancreas present in the human body produces insulin, a hormone that is responsible to help glucose reach each cell of the body as a source of energy for the body and its cells .Diabetes occurs when there are insufficient amounts of insulin or the body cells cannot use insulin well. Historically diabetes is a very old disease. In 1550 BC written on a 3rd Dynasty Egyptian papyrus, physician Hesyra mentions frequent urination as a symptom. Papyrus was purchased at Luxor in 1872s by Georg Ebers (A German archaeologist and writer). In this papyrus the symptoms of diabetes are described. The most extensive record of medical information and treatments were discovered on its translation. In India in 1500 BC polyuria in diabetes was associated with a sweet taste in Sanskrit texts of the 5th /6th century BCE, at the time of two notable physicians Sushruta and Charaka. They also identified the two types of diabetes mellitus, later dubbed type I and type II diabetes. Diabetes received its name from a Greek physician, Aretaeus of Cappadocia, after the word diabaikan which means "to siphon". [3]

There are three common types of diabetes; Type 1, Type 2 and Gestational. Type 1 occurs at a very young age of below 20 years when the pancreatic cells that produce insulin have been destroyed by the immune system. As a result, the pancreas does not make insulin at all. Type 2 happens when your body does not make or use insulin well and it is the most common type. Gestational tends to occur in pregnant women. In type 2 and gestational, pancreas produces insufficient amounts of insulin. All types of diabetes need to be treated and certainly early prediction will help to avoid them and their complications. The great progress that has happened in information technology and computer science will make the prediction easier. Knowledge discovery for predictive purposes is done through data mining, which is an analysis technique that helps in proposing inferences. This method helps in decision-making through algorithms from large amounts of data generated by these medical centers.[4]

First, this paper will discuss some information about diabetes including: (3) Types of diabetes, symptoms, common tests before suggesting some methodologies of prediction ending by a comparison between two of the highest accuracy methodologies.

Objectives:

The present work is intended to meet the following objectives:

1. **Give an introduction about Diabetes – types, medical tests, prevention, symptoms and diagnosis.**
2. **Show a brief introduction about machine learning techniques used for diabetes prediction.**
3. **Present a comparison between Random forest and SVM models for diabetes prediction.**
4. **Identify and discuss the field's benefits to the society along with effective application.**
5. **Future work.**

II. Acronyms

FPG: Fasting plasma glucose
BMI: body mass index
RPG: Random plasma glucose test
GHBA1C: Glycosylated hemoglobin
PPG: Post prandial Glucose
OGTT: Oral Glucose Tolerance Test
KNN: K nearest neighbor
SVM: Support vector machine
PPBS: Post prandial blood sugar

III. Overview of diabetes

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. Generally, a person is considered to be suffering from diabetes, when blood sugar levels are above normal (4.4 to 6.1 mmol/L). There are three main types of diabetes: Type 1, Type 2 and Gestational. [5]

Types of diabetes:

a. **Type 1:**

There are only about 10% of diabetes patients have this form of diabetes. In pancreas there are cells responsible for insulin production called beta cells. A patient with this type had his beta cells destroyed by the immune system at a very young age of below 20 years hence also called juvenile-onset diabetes. Unfortunately, the cause of type 1 diabetes is not known, and it is not preventable with current knowledge. However, its complications can be prevented or reduced by regulating blood sugar and taking insulin via pump or injection.

b. **Type 2:**

This type accounts for almost 90% of the diabetes cases and commonly called the adult-onset diabetes or non-insulin dependent

diabetes. It is associated with lifestyle habits. Obesity, poor diet and low activity are usually the causes of type 2.In this form, the various organs of the body become insulin resistant, and this increases the demand for insulin.

C. Gestational:

It is hyperglycemia with blood glucose values above normal but below those diagnostics of diabetes. It only occurs during pregnancy caused when increased hormones for the fetus leads to excess sugar in the blood and if mother's pancreas cannot produce enough insulin it leads to this form Most of the time, this type of diabetes goes away after the baby is born However, if you've had gestational diabetes, you have a greater chance o developing type2 diabetes later in life. Sometimes diabetes diagnose during pregnancy is actually type2 diabetes. As prevention, the patien must maintain a healthy diet and do regular blood testing within 24-2 weeks of gestational.

Generally, all these types are dangerous and need to be treated and periodically tested to avoid their fatal complications.

Symptoms:

There are common symptoms that appear on diabetics for example:

- **Polydipsia (excessive thirst)**
- **Weight gain or strange weight loss**
- **Healing of wounds is not quick, blurred vision, fatigue, itchy skin, etc.**
- **Polyuria (frequent urination)**
- **Sudden change in vision.**
- **Extremely dry skin and sores.**
- **Polyphagia (excessive hunger)**
- **Tingling or numbs in hands and feet.**

Complications and tests of diabetes:

- **Nerve damage (neuropathy).**
- **Kidney damage (nephropathy)**
- **Eye damage (retinopathy).**
- **Foot damage**
- **Skin conditions.**
- **Skin conditions.**
- **Alzheimer's disease.**
- **Depression. [6]**

Finally, the commonly conducted tests for determining whether a person has diabetes or not are:

RPG: (the normal range is from 70-160)

Sometimes health care professionals use the RPG test to diagnose diabetes when diabetes symptoms are present and they do not want to wait until you have fasted. You do not need to fast overnight for the RPG test. You may have this blood test at any time.

FPG: (the normal range is from 70-110)

The FPG blood test measures your blood glucose level at a single point in time. For the most

Reliable results, it is best to have this test in the morning, after you fast for 10-16 hours.

Fasting means having nothing to eat or drink except sips of water.

GHBA1C :(the normal range is from 4.5%-7%)

The A1C test is a blood test that provides your average levels of blood glucose over the past 3 months. Other names for the A1C test are hemoglobin A1C, HbA1C, glycated hemoglobin, and glycosylated hemoglobin test. You can eat and drink before this test. When it comes to using the A1C to diagnose diabetes, your doctor will consider factors such as your age and whether you have anemia NIH external link or another problem with your blood.1 The A1C test is not accurate in people with anemia.

OGTT: (The normal range for fasting: 100 to 140 mg/dL (3.3 to 5.5 mmol/L))

The OGTT measures blood glucose after you fast for at least 8 hours. First, a health care professional will draw your blood. Then you will drink the liquid containing glucose. For diagnosing gestational diabetes, you will need your blood drawn every hour for 2 to 3 hours.

PPBS: (the normal range is from 80-140)

A postprandial glucose (PPG) test is a blood glucose test that determines the amount of glucose, in the plasma after a meal. The diagnosis is typically restricted to postprandial hyperglycaemia due to lack of strong evidence of co-relation with a diagnosis of diabetes. But the American Diabetes Association do not recommend a PPG test for determining diabetes. [7]

IV. Why machine learning?

All these tests will cost the patient a lot especially if there are many people who are being tested. Technology and machine learning introduce an easy and fast way to predict diabetes using smaller number of tests and fixed parameters with an accuracy approaches around 80%.If we take an example: The Egyptian government held a campaign called "100 Million Health" to inventory the number of people who are diabetic or predicted to be diabetic in the future. With the help of machine learning we can make an easy and quick tests for all people and exclude all the people who are not possible to catch the disease and then do actual tests on people who have the probability to catch the disease to verify 100% if they are diabetic or not. So instead of doing tests on 100 million people we would have to check let say 30 million people. Although the doctors have their methods to predict the disease, we believe as engineers that we should be a part of diabetes prediction by our algorithms.

V. The relation between Machine learning and Linear algebra

Recently a lot of machine learning concepts are tied to linear algebra concepts. We have a huge number of attributes and data that need to be translated into numbers. This means that instead of dealing with just things or attributes, we represent these attributes (after translating them into numbers) in a set of vectors and matrices then we can do math on them. Machine learning applications are usually multidimensional applications and representing the data as points in a vector space is convenient. It's easy to map attributes of the data to dimensions in the vector space .Some basic examples, Principal component analysis(PCA)-eigenvalue, regression- matrix multiplication.....To do all these basic and important steps in machine learning we need to understand some concepts in linear algebra.[8]

VI. Related work

We agreed that the early detection or prediction of diabetes will help to avoid the disease and its complications. With the help of the recorded medical tests and the great development in the field of information technology and data mining it is possible to predict diabetes with accuracy practically around 80%.

Machine learning is useful when there is a large amount of example data and when the rules for prediction are unclear so we can use any machine learning algorithm such as k nearest neighbor (KNN), Decision Tree, Random forest, support vector machine (SVM) and neural networks.

We will use some patient data basement and divide it into two phases; the learning phase and the testing phase respectively considering that

the numbers of data used in the learning phase is greater than the number of testing phase.
Briefly we will compare them and show why we chose our algorithm.

a. KNN:

This algorithm represents the data in 2-D graph such that each point has an output result 0 or 1 after that it compares the points around the input based on k (the number of points that will be taken in the decision) to decide the output considering that its highest accuracy on test set is around: 0.78 when k=9.

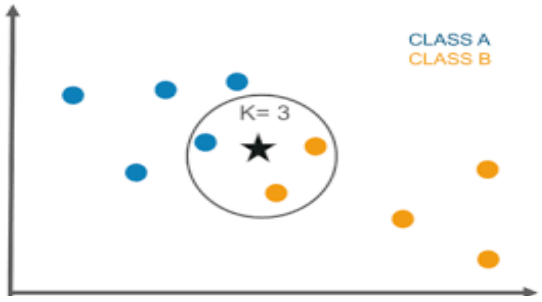


Figure 1_represent the number of neighborhood

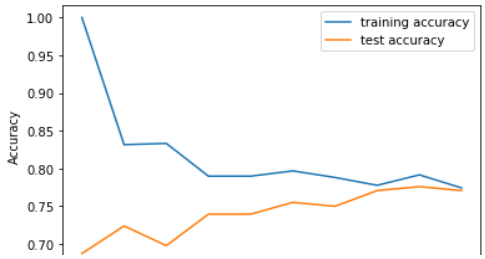


Figure 2_represent accuracy

b. Decision Tree:

Decision tree use the data set to make the program ready to know an unknown attribute. Decision Tree is supervised learning algorithm by using the training data and doing some algorithms on them. The accuracy of this method depends on the training data set and the size of the tree (how many categories I have?!). Its structure begins from the top to down, which are (a) Root node the first separation question in the model, (b) Internal node: attributes that locate on the inner part of the tree, (c) Branches descending of a node: The branch initiates a possible values for the attribute, (d) Leaf nodes: the predefined classes. Its accuracy on test set is around: 74%. [10]

c. Bagging:

Bagging (bootstrap aggregating) is a well-known ensemble method introduced by Leo Breiman. It aims to increase accuracy by making multiple versions of a predictor. To use training data set perfectly make a random draw with the replacement of examples. The outputs of the models are combined to create a single output. Bagging at many times produces a combined model that it performs better than the single model that produces from the single data.[10]

VII. Comparison between two of the highest accuracy algorithms

1. SVM (support vector machine)

Support-vector machine is one of the most important and famous machine learning algorithms are supervising models with associated learning techniques that analyses data which we need it for classification and regression analysis . if we have a set of training samples , each one of them marked as belonging to one or the other of two categories , an SVM training algorithm builds a model that assigns new samples to one category or the other , making it a non-probabilistic binary linear classifier

(Although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). It works on smaller datasets, but on them, it might be much stronger and powerful in building models. [13][14]

a. What space vector machine is?

An algorithm which we use for classification or regression processes, it is usually used for classification problems.

First, we plot each data item as a point in n-dimensional space (where n is number of features we have) with the value of each feature being the value of a particular coordinate. Then, we make classification by finding the '**hyper-plane**' that differentiates between the two classes as shown. If we use 2d plane and 2(as example): As appear in figure (3): there are a set of small red circles and another set for green stars. The **hyper-plane** in this case is line between the two classes.

b. How it works?

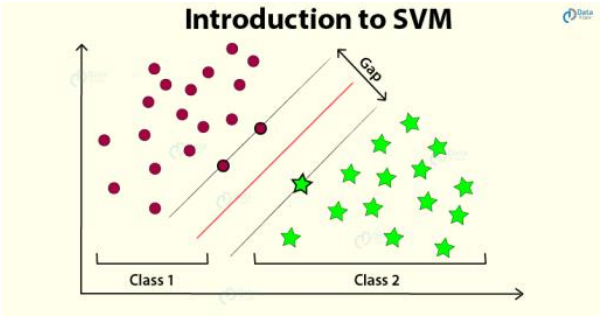


Figure 3

SVM stated finding the right hyper-plane we will learn how SVM plot this line:

So in the beginning the hyper-line is the line which separates and classifies between certain two categories, the area above the hyper-line is called +1 area ” class 1” and the area below the hyper-line is called -1 area “class 2”.

So, where we can put this line? Line A,B or C?

You should remember the definition of the hyper-plane which tries to differentiate the two classes the best way. In this scenario, hyper-plane "B" has excellently performed this job.

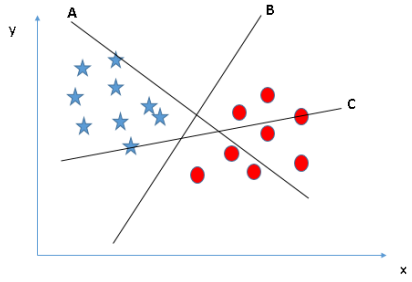


Figure 4

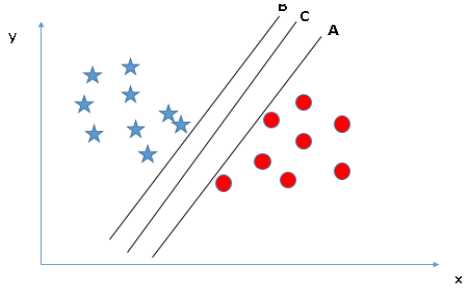


Figure 5

In **figure (5),** we find that 3 hyper-lines which differentiate the two classes better. You can get that the margin for hyper-line C is in the middle as compared to both A and B. Hence, we name the right hyper-line as C. Another important reason we can find for selecting.

The hyper-line with higher margin is robustness. If we select a hyper-line like A or B having low margin then there is high chance of Wrong classification.

In general we bisects the vacuum between two 2 classes. Or we plot the support vectors that touch the nearest element of the 2 classes and the hyper-line becomes the optimal one as shown in figure (6).

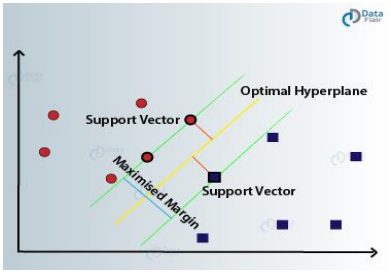


Figure 6

It is easy to have a linear hyper-plane between these two classes. But, another impportant question which arises is, should we need to add this feature manually to have a hyper-plane.

Answer is No, SVM has a technique called the kernel trick method. it converts not separable problem to separable problem. It is mostly useful in non-linear separation problem. it does some extremely complex data transformations, then find out the process to separate the data based on the labels or outputs you've defined, as shown in figure (7).

SVM has can also ignore outliers and find the hyper-line that has

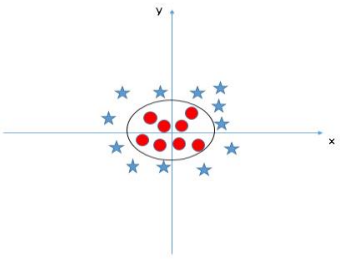


Figure 7

maximum m utliers. So, in figure (8),we find ' the below of the red small circles.

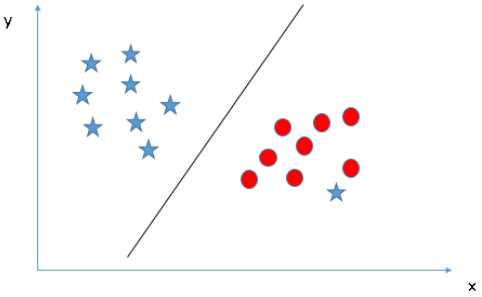


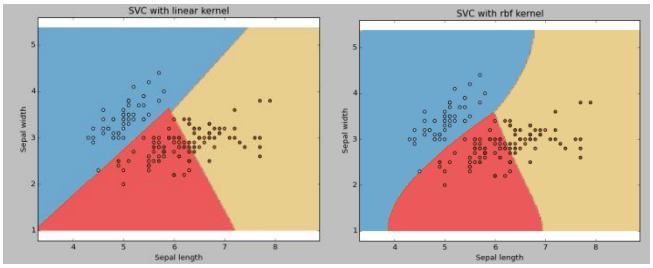
Figure 8

After classifying the input samples and separating between them by the hyper plane to two element and we get new element to test SVM plot this in the plane and if the new elements is above the hyper-plane so it is belong to the class 1 (the class above the hyper-plane), and if the new element is below the hyper-plane so it is belong to the class 2 (the class below the hyper-plane).

c. How to Tune SVM Parameters?

Kernel:

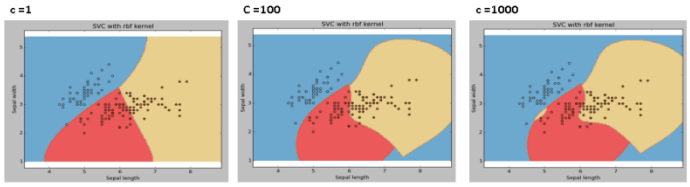
Kernel in the SVM is responsible for transforming the input data into the required format. Many of the kernels used in SVM are linear, polynomial and radial. For creating a non-linear hyper-plane,



we use Polynomial function, for complex applications. As shown in the circle in the in figure shown. [16]

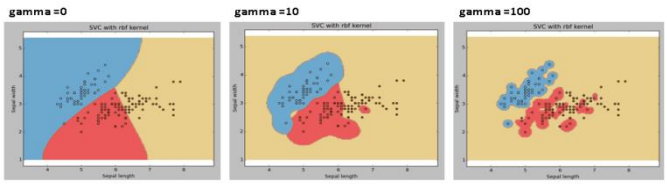
Regularization:

We can maintain regularization by adjusting it in the Scikit-learn's C parameters. C denotes a penalty parameter representing an error or any form of misclassification. With this misclassification, one can understand how much of the error is bearable. Through this, you can nullify the compensation between the misclassified term and the decision boundary. With a smaller C value, we obtain hyper-plane of small margin and with a larger C value; we obtain hyper-plane of larger value.



Gamma:

With a lower value of Gamma will create a loose fit of the training dataset. On the contrary, a high value of gamma will allow the model to get fit more appropriately. A low value of gamma only provides consideration to the nearby points for the calculation of a separate plane whereas the high value of gamma will consider all the data-points to calculate the final separation line.



d. **SVM equations:**

Linear SVM:

The symmetric line in the first fig 'A' which is called the hyper line separates between two categories... The area above the hyper line is called +1 area and the area below is called -1 area.

Let we have a vector (W) perpendicular to the hyper line and a vector (X) which is in the direction of the hyper line and we can analyze it into 2 components in x and y directions .. So Xi is the real vector indicated the points in the plane which includes the samples...

>> Hyper line (WX – b = 0)

>> Line which touch the first near sample above the hyper line or it is the line above hyper line to make the distance between them as large as possible (WX – b = 1)

>> Line which touch the first near sample below the hyper line or it is the line below hyper line to make the distance between them as large as possible (WX – b = -1)

>> The distance between the two lines above and below the hyper line equals (2/norm W)

>> to maximize this distance we want to minimize norm W by distance from a point to plane equations (WX – b >= 1 if Yi =1) , (WX – b <= -1 if Yi = -1)

These constraints state that each data point must lie on the correct side of the margin.

This can be rewritten as:

Yi (WX – b) >= 1 for all 1<= i <= n

Where n is the given point. [17], [18]

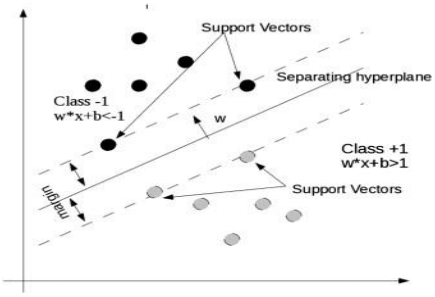


Figure 9

e. **Applications of SVM:**

When we have a dataset SVM can study this data and analyze and predict some anther data so we can use support vector machine in:

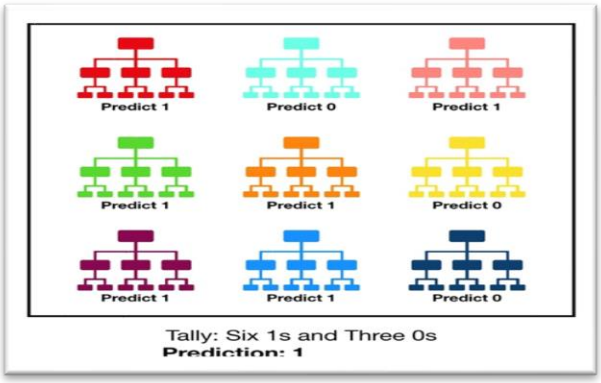
1. Diabetes prediction (it can predict with accuracy reaches 75 %.)
2. Handwriting recognition.
3. Face recognition.
4. Detecting steganography in digital
5. Breast cancer diagnosis

2. Random forest:

It is an updated method from the decision tree. Although Decision tree is simple to understand, interpret and visualize, it has some disadvantages that made it necessary to be improved. The most prominent disadvantage is that it cannot guarantee to return the globally optimal decision this can be mitigated by training multiple trees and form a forest. This leads us to use random forest method that depends on ensemble classification.

"Ensemble classification is an application of ensemble learning to boost the accuracy of classification. Ensemble learning is a machine learning paradigm where multiple models are used to solve the same problem. In ensemble classification, multiple classifiers are used and are more accurate than the individual classifiers in the ensemble. A voting scheme is then used to determine the class label for unlabeled instances. A simple and yet effective voting scheme is majority voting. In majority voting, each classifier in the ensemble is asked to predict the class label of the instance being considered. Once all the classifiers have been queried, the class that receives the greatest number of votes is returned as the final decision of the ensemble."[8]

This algorithm is consisting of creating many decision trees and the trees are arranged randomly then take the output of all individual



decision tree and it takes the majority as an output considering that its practical accuracy on test set is around 79.2%.

a. Mathematical background of random forest

For using random forest the code must calculate some important equations to build the forest and because that most of them are related to the decision tree such as

1- Gini impurity

Used by the CART (classification and regression tree) algorithm for classification trees, Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labelled if

Figure 10_Random forest example

the subset. The Gini impurity can be computed by summing the probability pi of an item with label i being chosen times the probability ($\sum_{k\neq 0}\ p_k = 1 - p_i$) of a mistake in categorizing that item. It reaches its minimum (zero) when all cases in the node fall into a single target category.

2- Variance reduction

Introduced in CART, variance reduction is often employed in cases where the target variable is continuous (regression tree), meaning that use of many other metrics would first require discretization before being applied. The variance reduction of a node *N* is defined as the total reduction of the variance of the target variable *x* due to the split at this node:

$$I_v(N) = \frac{1}{|s|^2}\sum_{i\epsilon s}\sum_{j\epsilon s}\frac{1}{2}(x_i - x_j)^2 - (\frac{1}{|s_t|^2}\sum_{i\epsilon s}\sum_{j\epsilon s}\frac{1}{2}(x_i - x_j)^2 + \frac{1}{|s_f|^2}\sum_{i\epsilon s}\sum_{j\epsilon s}\frac{1}{2}(x_i - x_j)^2)$$

where s, s_t, and s_f are the set of presplit sample indices, set of sample indices for which the split test is true, and set of sample indices for which the split test is false, respectively. Each of the above summands is indeed variance estimates, though, written in a form without directly referring to the mean.

b. How random forest algorithm work?

We discussed that random forest method is an updated method from the decision tree and in this section; we discuss how they are related mathematically and how Linear algebra and matrices help in this purpose.

In data mining, decision trees and random forest can be described also as the combination of mathematical and computational techniques to aid the description, categorization and generalization of a given set of data.

Data comes in records of the form:

$$(X, Y) = (X1, X2, X3, \dots\dots\dots\dots Xk, Y)$$

The dependent variable, Y, is the target variable that we are trying to understand, classify or generalize. The vector x is composed of the features $X1, X2, X3$…… etc., that are used for that task.

And if we try to reflect this form on our research:

$$(X, Y) = (pregnancy, Glucose, Blood\ pressure, \dots\dots\dots\dots, \text{Outcome})$$

The used data set will form a huge matrix of training samples that will be submitted to the algorithm to create a classification model. Assume that the data set form a matrix S with elements fij represent the features values.

Where: $i = A, B, C \dots\dots$; $J = 1,2,3, \dots\dots\dots\dots\dots N$

$$S = \begin{bmatrix} f_{A1} & f_{B1} & f_{C1} & \cdots & C_1 \\ & \vdots & & \ddots & \vdots \\ f_{A1} & f_{B1} & f_{C1} & \cdots & C_N \end{bmatrix}$$

The first column represent feature A of the sample and we continue the sample up to N samples. The same happens with the other features.

From this sample set we create a lot of subsets with random values, for example:

$$S_1 = \begin{bmatrix} f_{A12} & f_{B12} & f_{C12} & \cdots & C_{12} \\ & \vdots & & \ddots & \vdots \\ f_{A35} & f_{B35} & f_{C35} & \cdots & C_{35} \end{bmatrix}$$

$$S_2 = \begin{bmatrix} f_{A2} & f_{B2} & f_{C2} & \cdots & C_2 \\ & \vdots & & \ddots & \vdots \\ f_{A20} & f_{B20} & f_{C20} & \cdots & C_{20} \end{bmatrix}$$

$$S_M = \begin{bmatrix} f_{A4} & f_{B4} & f_{C4} & \cdots & C_4 \\ & \vdots & & \ddots & \vdots \\ f_{A12} & f_{B12} & f_{C12} & \cdots & C_{12} \end{bmatrix}$$

From the first subset S1 we create a decision tree number 1 then we make another random subset with different values to form the decision tree 2. In the same way we can form decision trees up to M. This is why we call this algorithm Random forest (lots of decision trees).

With all of these decision trees we have different variations of the main classification. Then with the help of these decision trees we can create a ranking of classifiers and if we have a new element to classify we are going to ask each tree for its prediction (If the tested person is diabetic or not) and by accounting for the number of votes for each classifier we get a final prediction.

c. **Measurement:**

In this study, we used sensitivity (SN), specificity (SP), accuracy (ACC), and Matthews's correlation coefficient (MCC) to measure the classified effectiveness. And the formulas are as follow:

$$SN = \frac{TP}{TP + FN}$$

$$SP = \frac{TN}{TN + FP}$$

$$ACC = \frac{TN + TP}{TN + TP + FP + FN}$$

$$= \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}$$

Which the true positive represents (TP) the number of identified positive samples in the positive set. True negative (TN) means the number of classification negative samples in the negative set. False positive (FP) is the number of identified positive samples in the negative set. And false negative (FN) represents the number of identified negative samples in the positive set. The accuracy is defined as the ratio of the number of samples correctly classified by the classifier to the total number of samples. In medical statistics, there are two basic characteristics, sensitivity (SN) and specificity (SP). Sensitivity is the true positive rate, and specificity is the true negative rate. The MCC is a correlation coefficient between the actual classification and the predicted classification. Its value range is [-1, 1]. When the MCC equals one, it indicates a perfect prediction for the subject. When the MCC value is 0, it indicates the predicted result is not as good as the result of random prediction, and -1 means that the predicted classification is completely inconsistent with the actual classification".[11]

d. **Model Evaluation**

Evaluation is the processes to calculate the effectiveness of the results for data analysis models. Accuracy of the performance of a classification model is based on the count of the test records correctly and incorrectly predicted by the model. These counts are tabulated in a table known as a confusion matrix. In the case of two classes, the Accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. Accuracy is a popular evaluation performance of a classifier. Most classification algorithms seek models that attain the highest accuracy when applied to the test set. [9]

VIII. The algorithm that is used and the reason

We used random forest model for further optimization because we found that a random forest model may be easier to study and analyze compared to the "black box" SVM model which are the two highest accuracy algorithms. As we attempted to optimize the models, we found that when the model was optimized for high sensitivity, specificity was reduced, and vice versa. A model which would predict most hypoglycemia but would produce many false positives would not be able to deliver meaningful interventions to patients; this is what unfortunately happens in SVM. The opposite happens in random forest where it has an error of giving true negatives. This means that the error in random forest will predict a good person, who is not probably diabetic, as diabetic so he must do some actual tests to verify. This is better than telling some that he is good, and he has a great chance to get diabetes. So, we create a model that was optimized for a good balance between sensitivity and specificity using Random Forest method. [20]

IX. Results and Discussion (Obtained Using Google Colab (Python language))

In our experimental work we preferred using 8 parameters and they are Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function and Age. We have major parameters that include the authority to judge if the tested person is diabetic or not, and they are; Glucose and Diabetes pedigree function. The others are also important but partially depend on the major parameters such as; BMI, Age and pregnancies.

We use data basement with dimensions 768×9 (PIMA data set), five hundred people are not diabetic, and the rest are diabetic.

Each parameter in the data set is represented in a range of values which show the case of the tested person. But these values sometimes are not numerically related to the values obtained from laboratories because they are calculated from formulas that are not in the scope of this search.

All the results obtained by Google Colab (Python language)		
	Range of possible values	
Parameters	From	To
pregnancies	0	17
Glucose	0	200
Blood pressure	0	122
Skin thickness	0	99
insulin	0	846
BMI $\left(\frac{weight}{(height)^2}\right)$	18.2	67.1
Diabetes pedigree function	0.078	2.42
Age	21	81

Table (1)

By starting with average values that lead to no diabetes then trying to change the parameters to identify to what extent the test depends on this parameter.

pregnancies	Glucose	Blood pressure	Skin thickness	insulin	BMI	Diabetes pedigree function	Age
4	120	70	21	80	32	0.47	33
Test Result	0						

Table (2) represents the average values

Glucose is the most important parameter. By exceeding certain limit, it leads directly to diabetes.

pregnancies	Glucose	Blood pressure	Skin thickness	insulin	BMI	Diabetes pedigree function	Age
4	128	70	21	80	32	0.47	33
Test Result	1						

Table (3) Glucose effect

The next parameter is diabetes pedigree function which plays an effective role but less important than the Glucose.
As noticed in table (4), we had to increase the glucose as it is the major parameter.

pregnancies	Glucose	Blood pressure	Skin thickness	insulin	BMI	Diabetes pedigree function	Age
4	125	70	21	80	32	0.7	33
Test Result	1						

Table (4) Diabetes pedigree function effect

Another important parameter is the Age which will be a decisive factor especially after the age of forty.

pregnancies	Glucose	Blood pressure	Skin thickness	insulin	BMI	Diabetes pedigree function	Age
4	128	70	21	80	32	0.47	40
Test result	1						

Table (5) Age effect

We do not need to discuss the other parameters as they are extensively related to the discussed ones. The importance of used parameters is varying around the values in figure (11):

As we see in figure (12): the accuracy of the KNN and decision tree

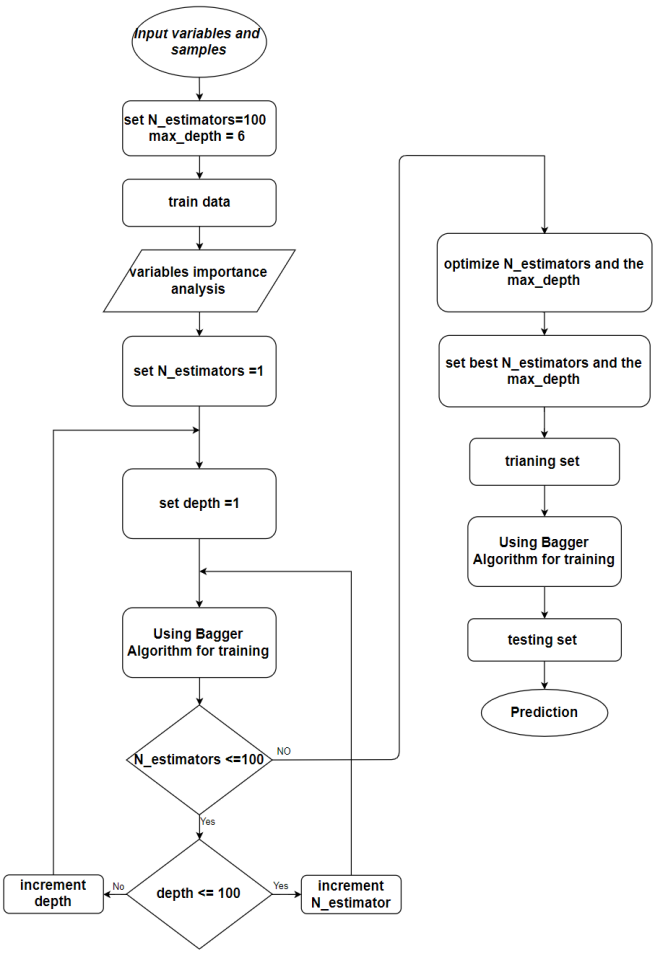
is very low compared to random forest and SVM , and for comparing between random forest and SVM we found that SVM have an error of making false positive and random forest gives only true negative. So, we decide that telling the patient that he has diabetes when he doesn't is better than telling him that he is well when he isn't.

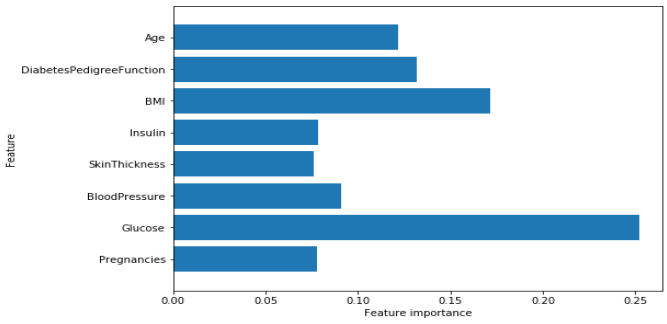
Figure 11 importance of each parameter

no of patient	1	2	3	4	5	6	7	8	9	10
Real tests	0	0	1	1	1	1	0	0	0	0
Random forest	0	1	1	1	1	1	0	0	0	0
SVM	0	0	0	1	1	1	0	0	0	0
KNN	0	0	1	1	1	0	0	0	1	1
Decision tree	1	0	1	1	0	1	0	0	0	1

The output of the four algorithms using random data
(for explanation only)

Figure 12


Input variables and samples
set N_estimators=100 max_depth = 6
train data
variables importance analysis
set N_estimators =1
set depth =1
Using Bagger Algorithm for training
N_estimators <=100
NO
Yes
depth <= 100
No
Yes
increment depth
increment N_estimator
optimize N_estimators and the max_depth
set best N_estimators and the max_depth
trianing set
Using Bagger Algorithm for training
testing set
Prediction


Age
DiabetesPedigreeFunction
BMI
Insulin
SkinThickness
BloodPressure
Glucose
Pregnancies
Feature
0.00
0.05
0.10
0.15
0.20
0.25
Feature importance