

# **DIABETES DETECTION IN WOMEN USING MACHINE LEARNING**

**A MINI PROJECT  
REPORT**

*Submitted by*

**SHRUTI RAJ VANSI SINGH  
ENTRY No. 17BCS049  
SANYA NEGI  
ENTRY No. 17BCS046**

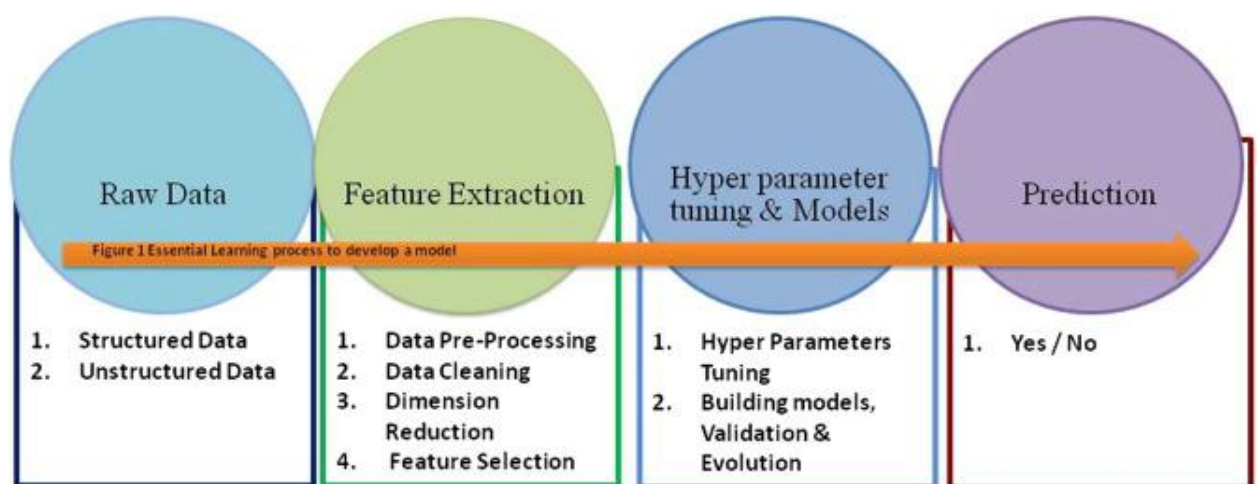
**BACHELOR OF TECHNOLOGY  
IN  
COMPUTER SCIENCE**



**SHRI MATA VAISHNO DEVI UNIVERSITY**

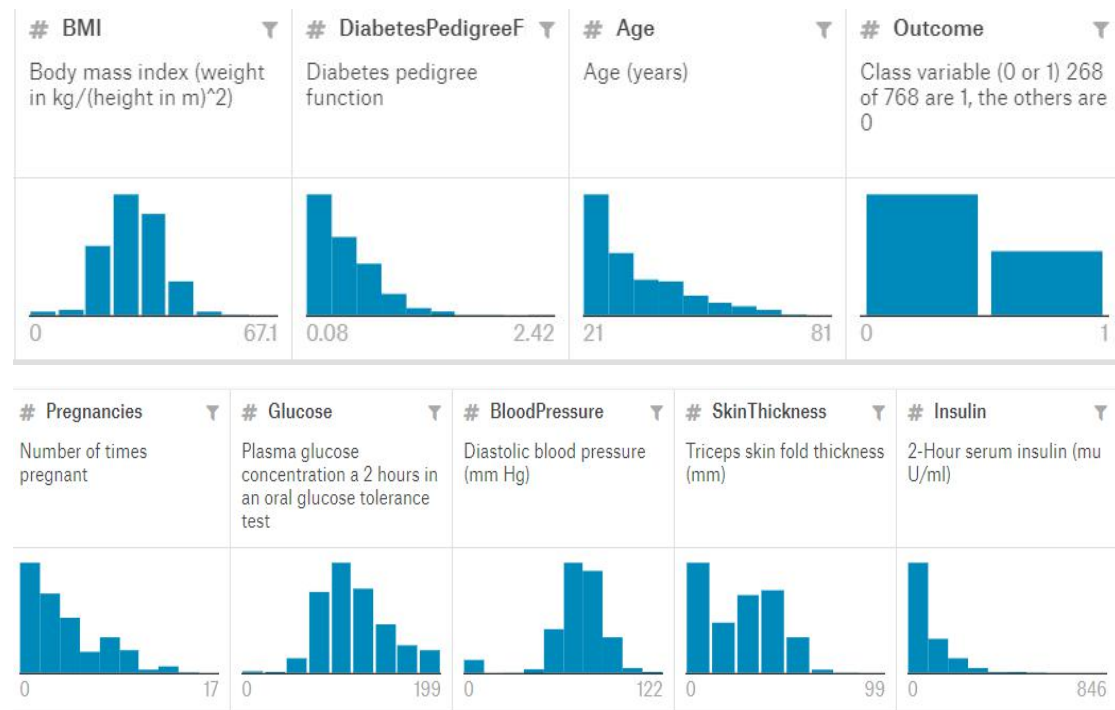
NOVEMBER 2019

The learning process starts with the gathering of data by different means, from various resources. Then the next step is to prepare the data, that is pre-process it in order to fix the data related issues and to reduce the dimensionality of the space by removing the irrelevant data (or selecting the data of interest). Since the amount of data that is being used for learning is large, it is difficult for the system to make decisions, so algorithms are designed using some logic, probability, statistics, control theory etc. to analyze the data and retrieve the knowledge from the past experiences. Next step is testing the model to calculate the accuracy and performance of the system. And finally optimization of the system, *i.e.* improvising the model by using new rules or data set. The techniques of machine learning are used for classification, prediction and pattern recognition. Machine learning can be applied in various areas like: search engine, web page ranking, email filtering, face tagging and recognizing, related advertisements, character recognition, gaming, robotics, disease prediction and traffic management. The essential learning process to develop a predictive model is given in Fig below.



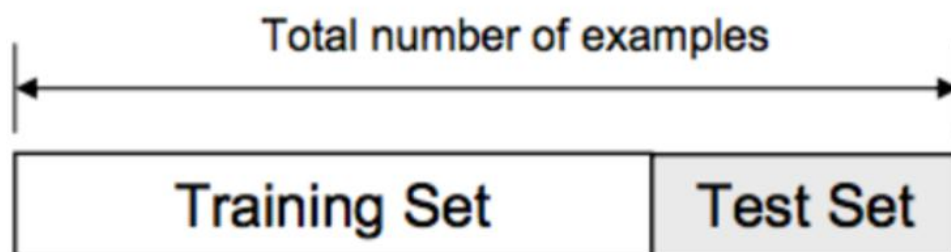
**Fig. 1.** Essential Learning process to develop a predictive model.

The following figure 2 is the graphical representation of various features helpful in detecting diabetes



**Fig 2.** Graphical representation

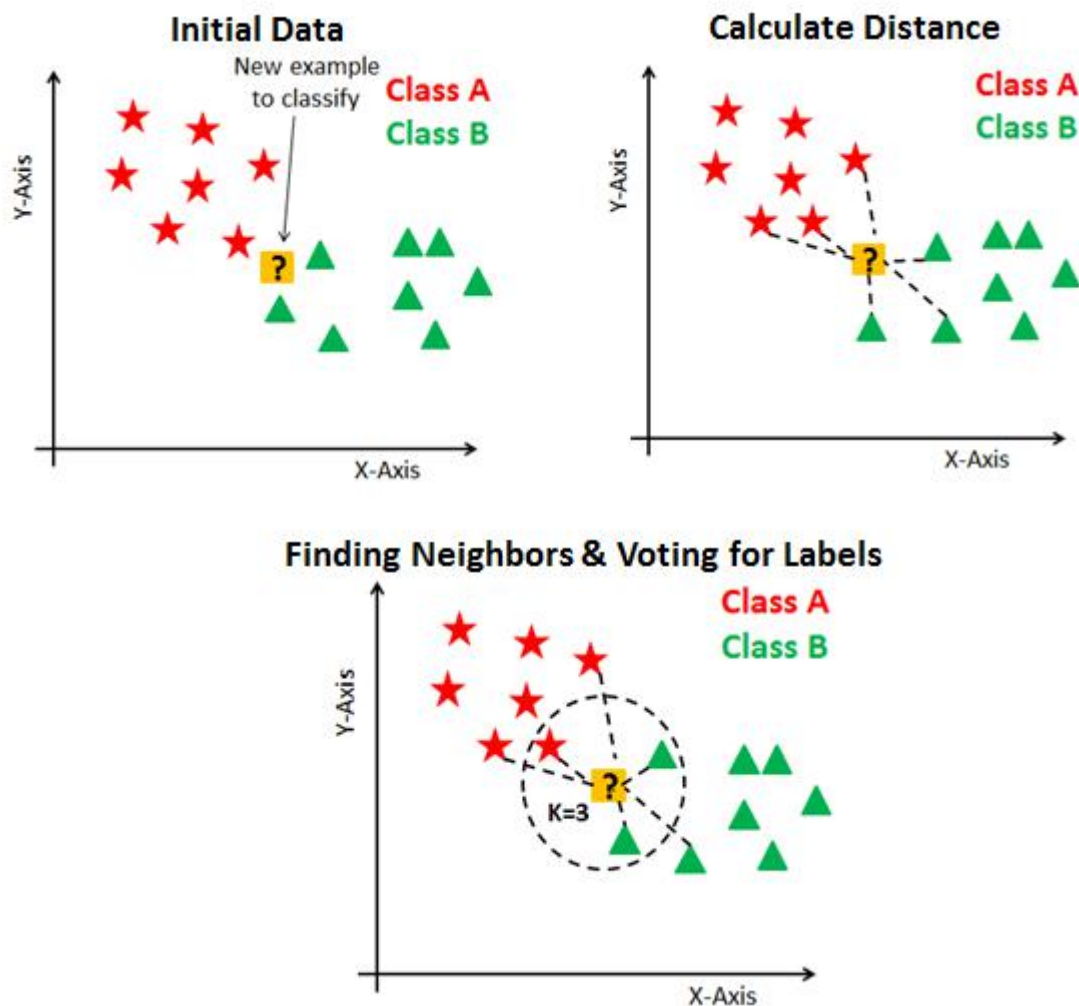
Categorical data are variables that contain label values rather than numeric values. The number of possible values is often limited to a fixed set. But in the given data set we do not need encode the categorical data.



**Fig. 3** Splitting of Data

For this, a ' $k$ ' is decided (where  $k$  is number of neighbours to be considered) which is generally odd and the distance between the data points that are nearest to the objects is calculated by the ways like Euclidean's distance, Hamming distance, Manhattan distance or Minkowski distance.

After calculating the distance, ' $k$ ' nearest neighbours are selected the resultant class of the new object is calculated on the basis of the votes of the neighbours. The  $k$ -NN predicts the outcome with high accuracy.



**Fig. 6**  $k$  Nearest Neighbour

Mathematically, kernel trick (K) is defined as:

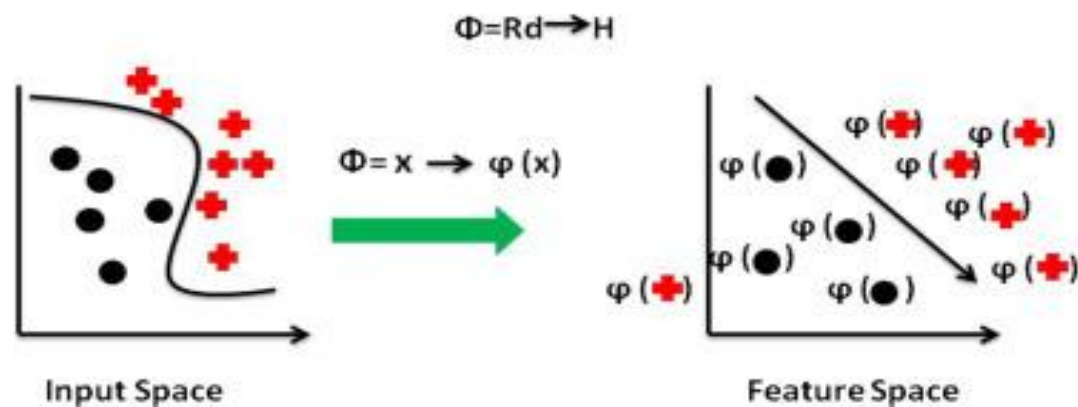
$$K(x_1, x_2) = \exp\left(-\frac{|x_1 - x_2|^2}{2\sigma^2}\right)$$

A Gaussian function is also known as Radial basis function (RBF) kernel.

In Figure, the input space separated by feature map ( $\Phi$ ). By applying equation (1), (2) we get:

$$f(\mathcal{X}) = \sum_i^N \alpha_i y_i k(\mathcal{X}_i, \mathcal{X}) + b$$

Function  $\Phi$  mapping the idea into another space is defined as:



**Fig. 8.** Representation of Radial basis function (RBF) kernel Support Vector Machine.

By applying equation (3) in 4 we get new function, where N represents the trained data.

$$f(\mathcal{X}) = \sum_i^N \alpha_i y_i \exp\left(-\frac{|x_1 - x_2|^2}{2\sigma^2}\right) + b$$

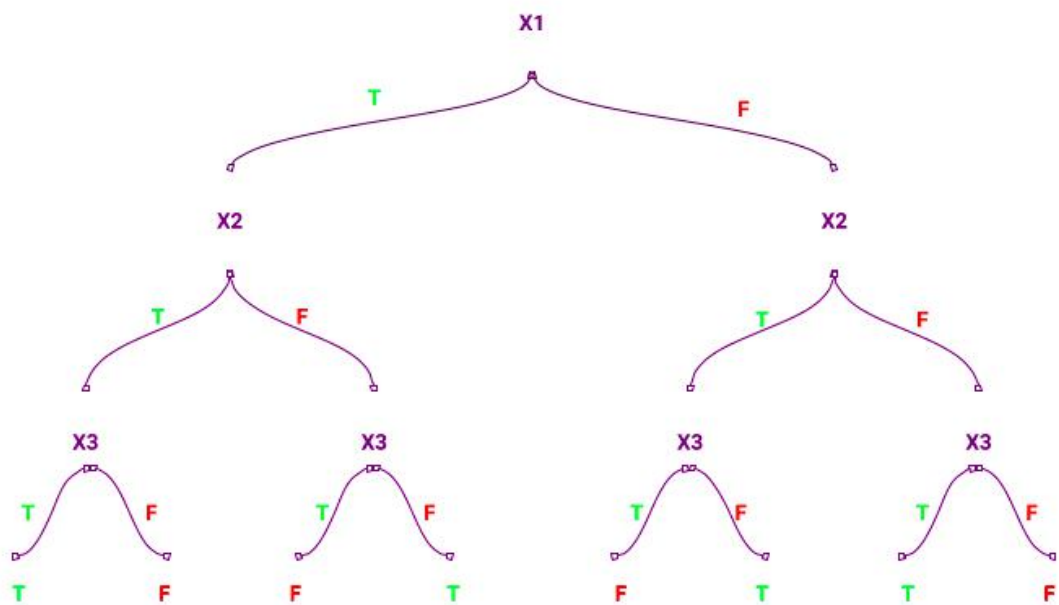
## 2.2.5 Naives Bayes Classification

Naive Bayes classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of

### 2.2.6 Decision Tree Classification

A decision tree is a tree-like graph with nodes representing the place where we pick an attribute and ask a question; edges represent the answers to the question; and the leaves represent the actual output or class label. They are used in non-linear decision making with simple linear decision surface.

Decision trees classify the examples by sorting them down the tree from the root to some leaf node, with the leaf node providing the classification to the example. Each node in the tree acts as a test case for some attribute, and each edge descending from that node corresponds to one of the possible answers to the test case. This process is recursive in nature and is repeated for every sub tree rooted at the new nodes.

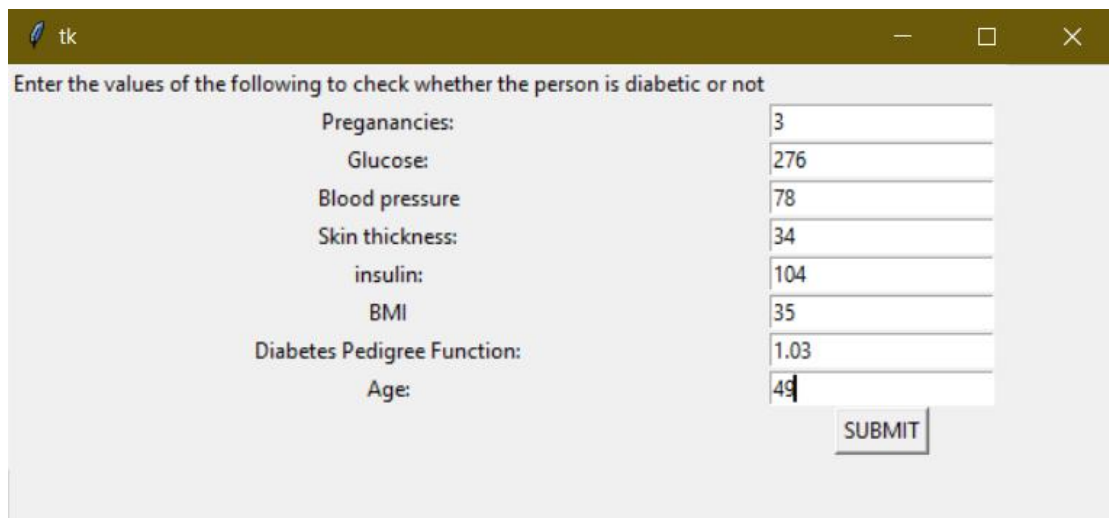


**Fig 9.** Decision tree for an XOR operation involving three operands

We used standard scaler from the sklearn preprocessing library to get all the values of features between -1 and 1 to make them comparable.

## 2.6 Graphical User Interface

We also built a GUI to support our model using tkinter in python. The GUI requires the user to enter the values of features to get the prediction whether the user is Diabetic or not.



Feature	Value
Preganancies:	3
Glucose:	276
Blood pressure	78
Skin thickness:	34
insulin:	104
BMI	35
Diabetes Pedigree Function:	1.03
Age:	49

**Fig 12.**Graphical User Interface

```
The Patient is Diabetic
In [8]:
```

**Fig 13.** Result for the data user entered

