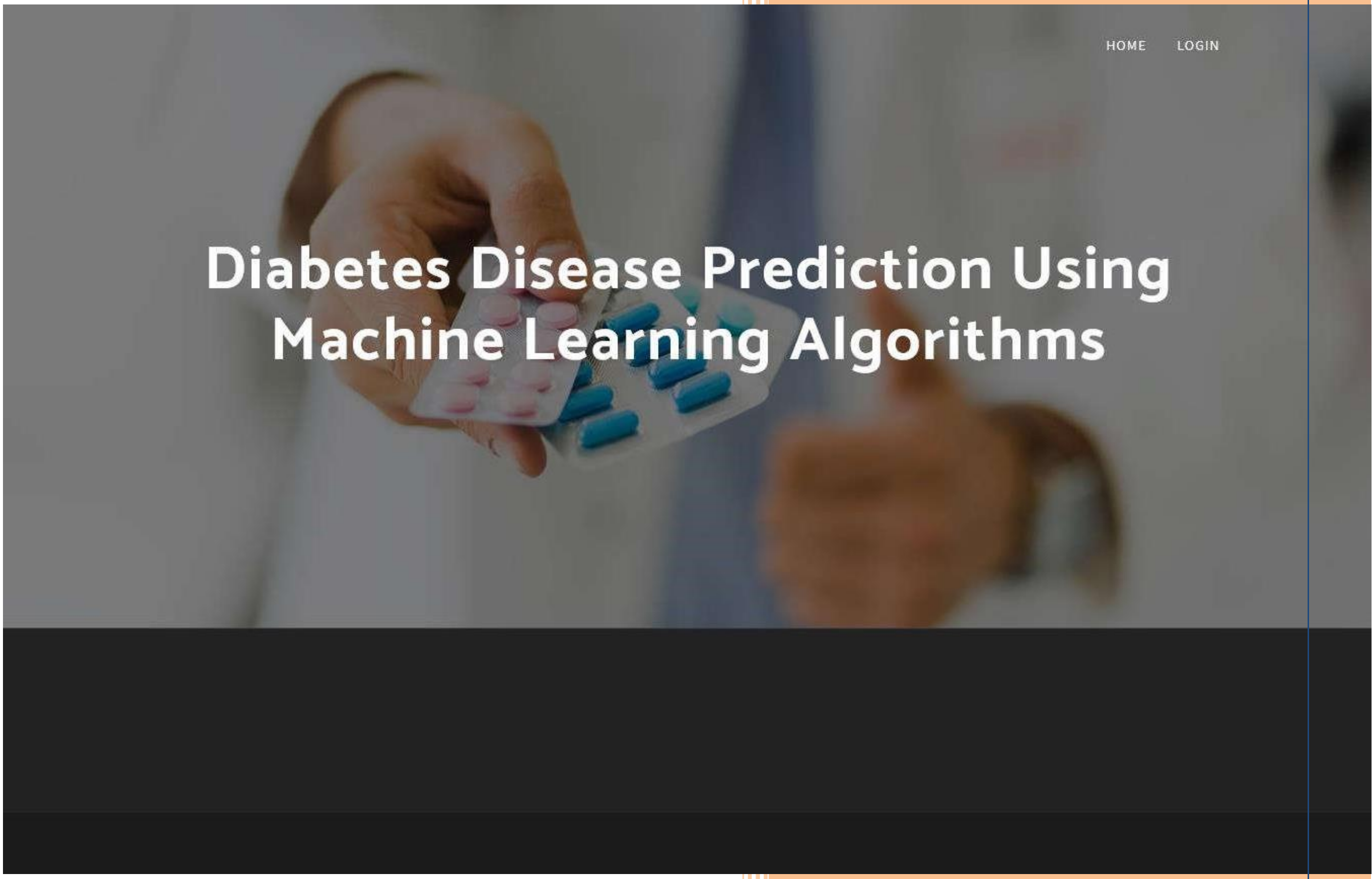
Diabetes Prediction Using Machine Learning Algorithms



HOME LOGIN

Diabetes Disease Prediction Using Machine Learning Algorithms

Group Members: -

1] Prit Mayani-20C21028

2] Smit Miyani-20C21031

3] Ved Patel-20C21056

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.

CONTENTS:

1. INTRODUCTION

All around there are numerous ceaseless infections that are boundless in evolved and developing nations. One of such sickness is diabetes. Diabetes is a metabolic issue that causes blood sugar by creating a significant measure of insulin in the human body or by producing 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 coronary failure, visual deficiency, kidney ailments and nerve harm, and so on.

Subsequently, the identification of such chronic metabolic ailment at a beginning period 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 various domains we may have the option to find an answer for this issue. ML strategies and neural systems help scientists to find new realities from existing well-being-related informational indexes, which may help in ailment supervision and detection. The current work is completed utilizing the Pima Indians Diabetes Database. The point of this framework is to make an ML model, which can anticipate with precision the likelihood or the odds of a patient being diabetic. The ordinary distinguishing process for the location of diabetes is that the patient needs to visit a symptomatic focus. One of the key issues of bio-informatics examination is to achieve precise outcomes from the information. Human mistakes or various laboratory tests can entangle the procedure of identification of the disease. This model can foresee whether the patient has diabetes or not, aiding specialists to ensure that the patient in need of clinical consideration can get it on schedule and also help anticipate the loss of human lives.

DNA makes neural networks the apparent choice. Neural networks use neurons to transmit data across various layers, with each node working on a different weighted parameter to help predict diabetes.

Presently, with the ascent of machine learning, AI, and neural systems, and their application in various domains we may have the option to find an answer for this issue. ML strategies and neural systems help scientists to find new realities from existing well- being-related informational indexes, which may help in ailment supervision and detection. The current work is completed utilizing the Pima Indians Diabetes Database.

Causes of Diabetes:

Genetic factors are the main cause of diabetes. It is caused by at least two mutant genes in the chromosome 6, the chromosome that affects the response of the body to various antigens. Viral infection may also influence the occurrence of type 1 and type 2 diabetes. Studies have shown that infection with viruses such as rubella, Coxsackie virus, mumps, hepatitis B virus, and cytomegalovirus increase the risk of developing diabetes.

Types of Diabetes:

Type 1-
Type 1 diabetes means that the immune system is compromised and the cells fail to produce insulin in sufficient amounts. There are no eloquent studies that prove the causes of type 1 diabetes and there are currently no known methods of prevention.

Type 2 -
Type 2 diabetes means that the cells produce a low quantity of insulin or the body can’t use the insulin correctly. This is the most common type of diabetes, thus affecting 90%mof persons diagnosed with diabetes. It is caused by both genetic factors and the manner of living.

2. DATASET

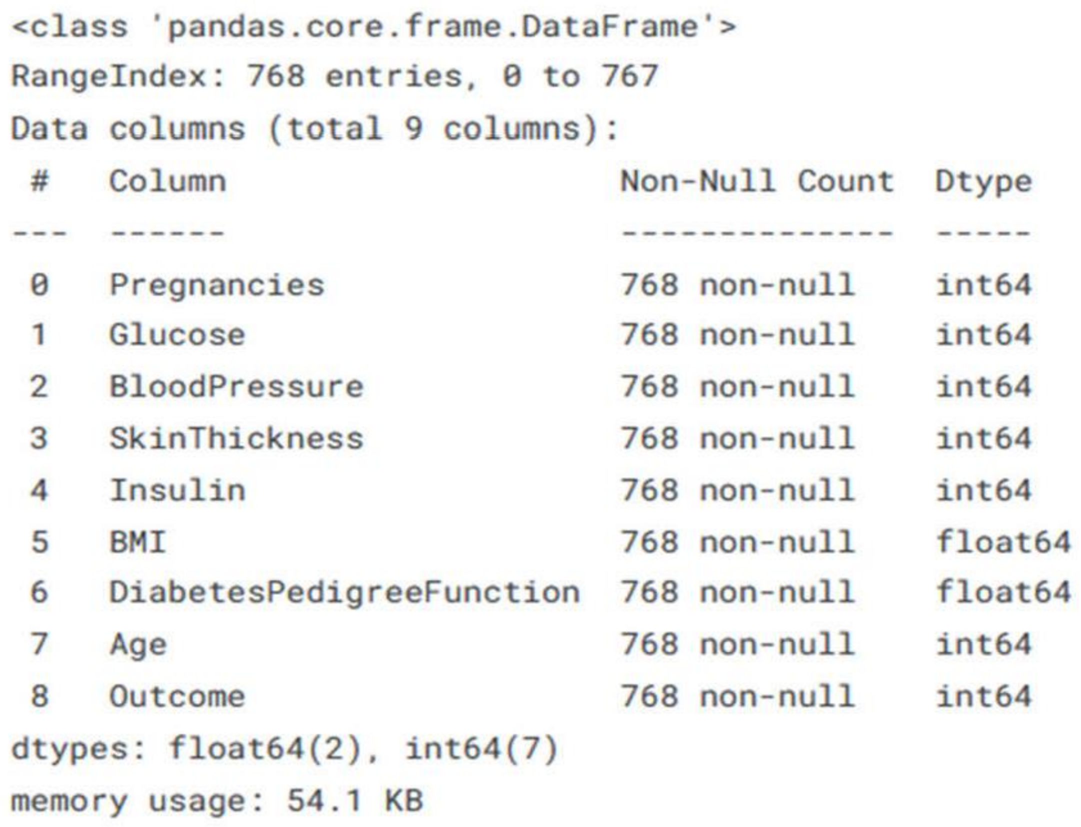
The dataset collected is originally from the Pima Indians Diabetes Database is available on Kaggle. It consists of several medical analyst variables and one target variable. The objective of the dataset is to predict whether the patient has diabetes or not. The dataset consists of several independent 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	1Pregnancies Number of times pregnant
2	Glucose	2 Glucose Plasma glucose concentration
3	Blood Pressure	3 Blood Pressure Diastolic blood pressure
4	Skin Thickness	4 Skin Thickness Triceps skin fold thickness (mm)
5	Insulin	5 Insulin 2-h serum insulin
6	BMI	6 BMI Body mass index
7	Diabetes pedigree function	7 Diabetes pedigree function Diabetes pedigree function
8	Outcome	8 Outcome Class variable (0 or 1)
9	Age	Age of patient

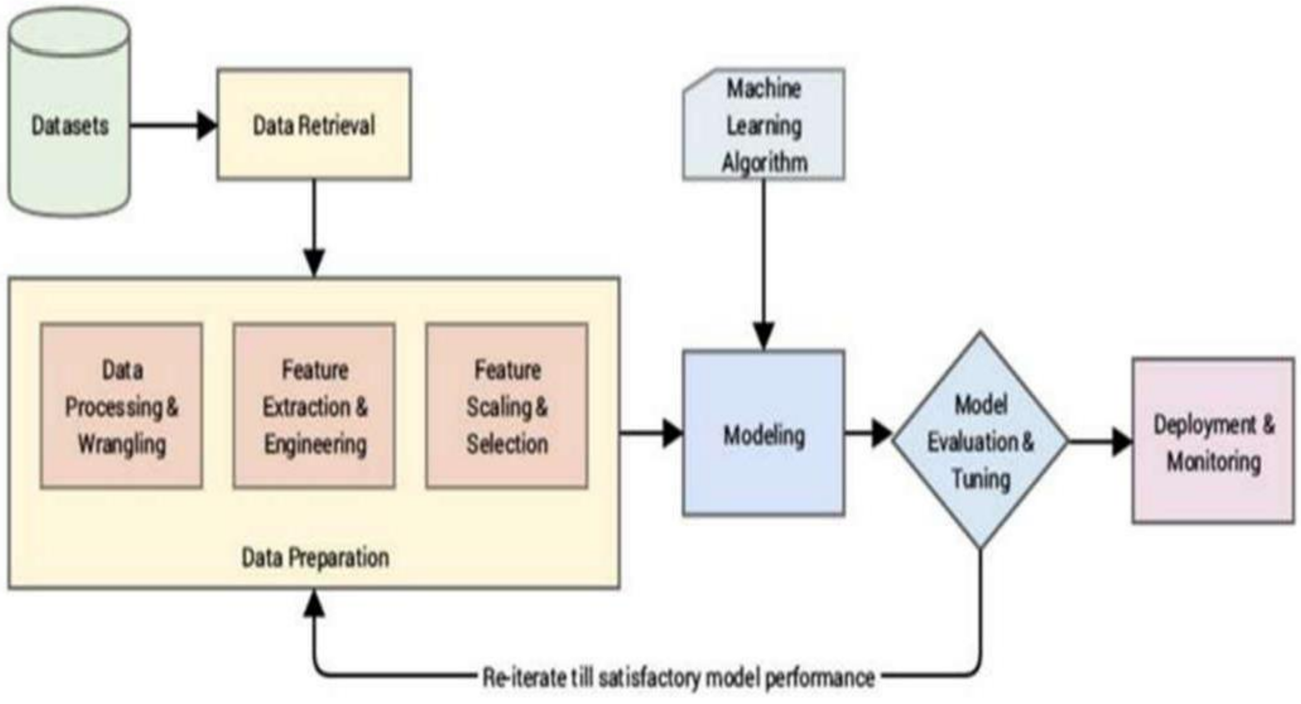
➔The diabetes data set consists of 2000 data points, with 9 features each.

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



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

➔There is no null values in dataset.


Datasets
Data Retrieval
Machine Learning Algorithm
Data Processing & Wrangling
Feature Extraction & Engineering
Feature Scaling & Selection
Data Preparation
Modeling
Model Evaluation & Tuning
Deployment & Monitoring
Re-iterate till satisfactory model performance

3. PROPOSED METHODS

I] Dataset collection:

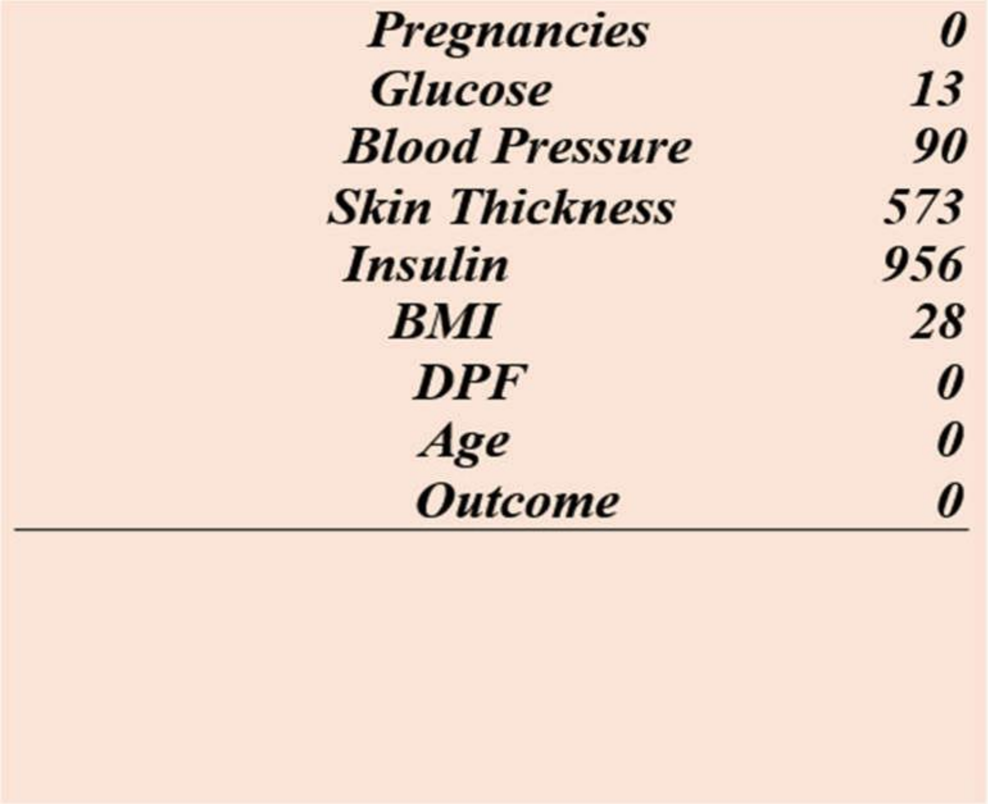
It includes data collection and understanding the data to study the hidden patterns and trends which helps to predict and evaluating the results. Dataset carries 1405 rows i.e., total number of data and 10 columns i.e., total number of features. Features include Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age.

II] Data Pre-processing:

This phase of model handles inconsistent data in order to get more accurate 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 selected attributes like Glucose level, Blood Pressure, Skin Thickness, BMI and Age because these attributes cannot have values zero. Then data was scaled using Standard Scaler. Since there were a smaller number of features and important for prediction so no feature selection was done.

III] Misplaced value identification:

Using the Panda library and SK-learn, we got the missing values in the datasets,. We replaced the missing value with the corresponding mean value

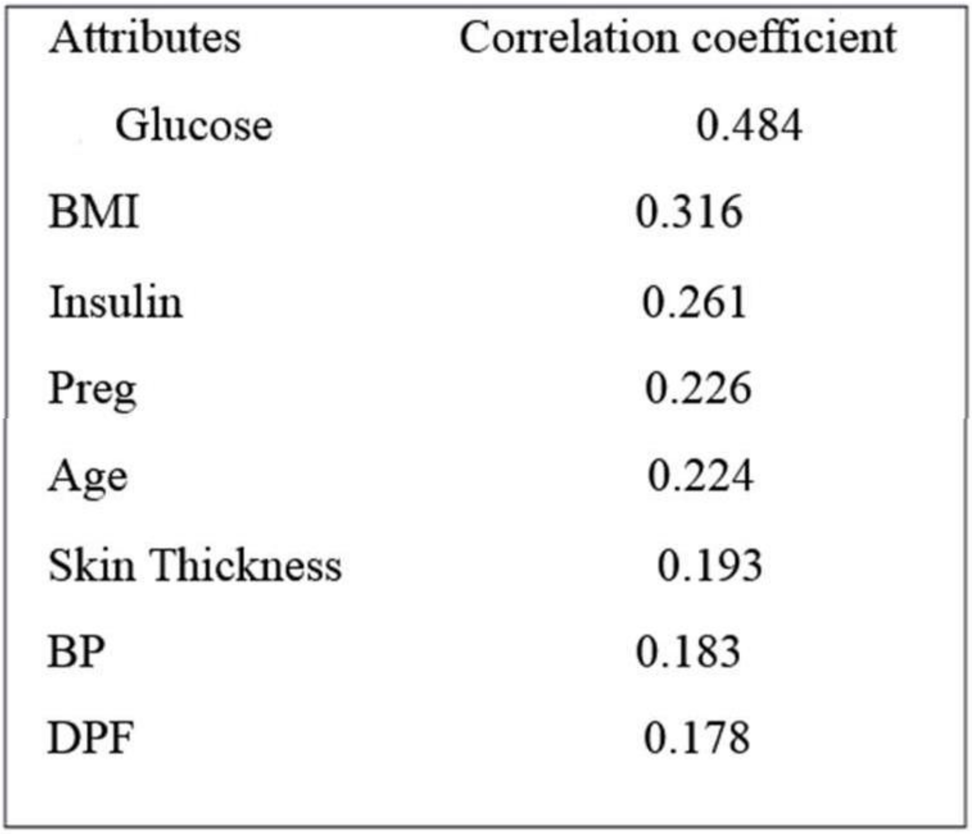


Pregnancies	0
Glucose	13
Blood Pressure	90
Skin Thickness	573
Insulin	956
BMI	28
DPF	0
Age	0
Outcome	0

Misplaced Data

IV] Feature selection:

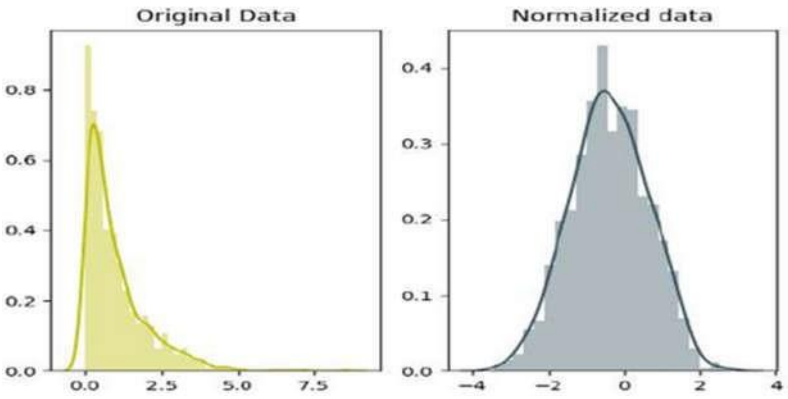
Pearson’s correlation method is a popular method to find the most relevant attributes/features. The correlation coefficient is calculated in this method, which correlates with the output and input attributes. The coefficient value remains in the range by between −1 and 1. The value above 0.5 and below −0.5 indicates a notable correlation, and the zero value means no correlation.



Attributes	Correlation coefficient
Glucose	0.484
BMI	0.316
Insulin	0.261
Preg	0.226
Age	0.224
Skin Thickness	0.193
BP	0.183
DPF	0.178

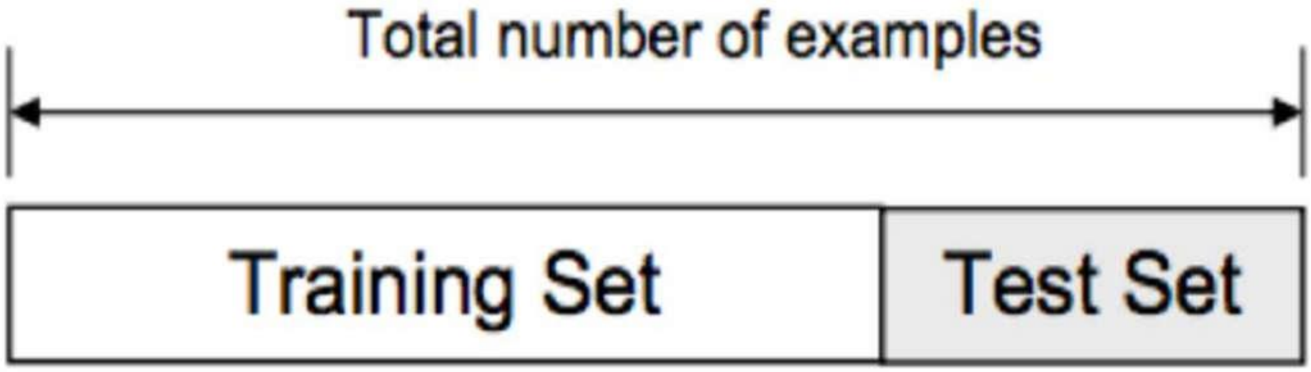
V] Scaling and Normalization:

We performed feature scaling by normalizing the data from 0 to 1 range, which boosted the algorithm’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 far apart 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] Splitting 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.



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 and ensemble techniques to predict diabetes dataset. We have applied SVM, LR, DT and RF Machine learning classifier to analyses the performance by finding accuracy of each classifier All the classifiers are implemented using scikit learn libraries in python. The implemented classification algorithms are described in next section.

4. MODELING AND ANALYSIS

A] K-Nearest Neighbors:

K-nearest neighbors (KNN) algorithm uses 'feature similarity' to predict the values of new data points 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.

B] 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.

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] 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 used for splitting the nodes.

E] Decision Tree:

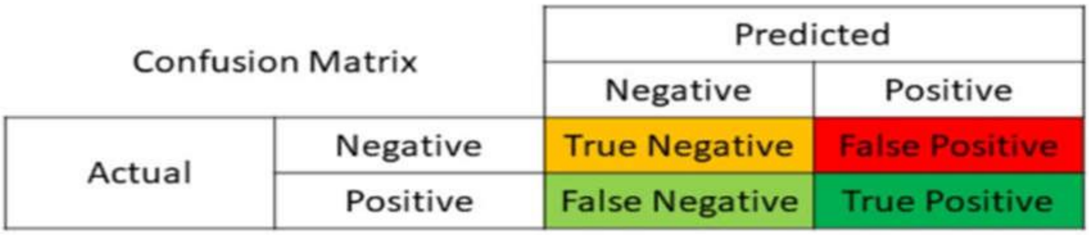
Decision tree is non parametric classifier in supervised learning. In this method all the details are represented in the form of tree, where leaves are corresponds to the class labels and attributes are corresponds to internal node of the tree. We have used Gini Index for splitting the nodes.

F] Logistic Regression:

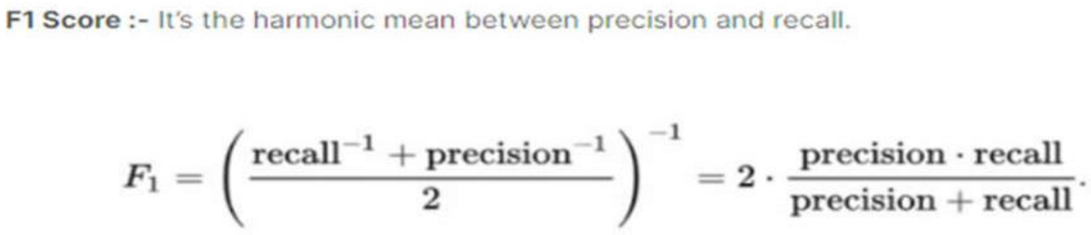
Logistic regression is a machine learning technique used when dependent variables are able to categorize .The outputs obtained by using the logistic regression is based on the available features. Here sigmoidal function is used to categorize the output.

5. MEASURNMENT

Performance Metrics :- Confusion Matrix, F1 Score, Precision Score, Recall Score **Confusion Matrix** It is a tabular visualization of the model predictions versus the ground-truth labels.

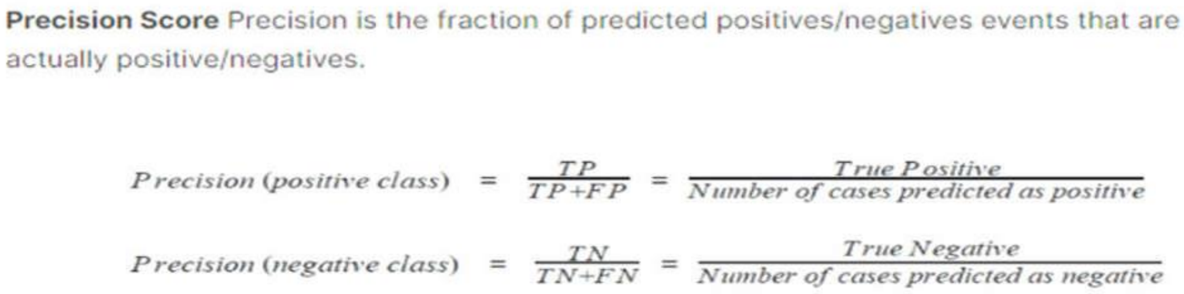


Confusion Matrix		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive



F1 Score :- It's the harmonic mean between precision and recall.

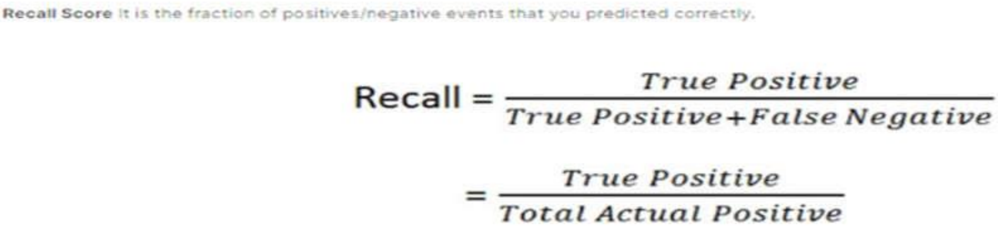
$$F_1 = \left(\frac{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}}{2}\right)^{-1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$



Precision Score Precision is the fraction of predicted positives/negatives events that are actually positive/negatives.

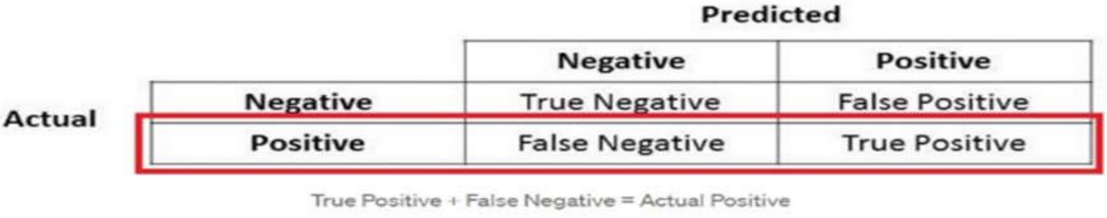
$$Precision\ (positive\ class) = \frac{TP}{TP+FP} = \frac{True\ Positive}{Number\ of\ cases\ predicted\ as\ positive}$$

$$Precision\ (negative\ class) = \frac{TN}{TN+FN} = \frac{True\ Negative}{Number\ of\ cases\ predicted\ as\ negative}$$



Recall Score It is the fraction of positives/negative events that you predicted correctly.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative} = \frac{True\ Positive}{Total\ Actual\ Positive}$$



		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

True Positive + False Negative = Actual Positive

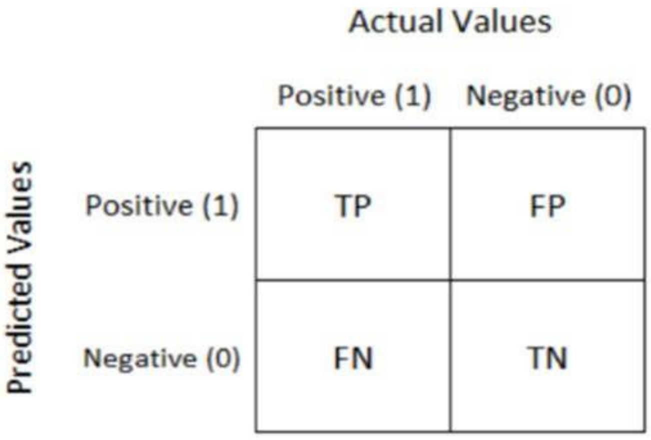
Confusion matrix: - which provides output matrix with complete description performance of the model. Here,

TP: True positive

FP: False positive

TN: True negative

FN: False negative



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

The following performance metrics are used to calculate the presentation of various algorithms.

- True positive (TP) – person has disease, and the prediction also has a positive.
- True negative (TN) – person not having disease and the prediction also has a negative.

- False positive (FP) – person not having disease but the prediction has a positive.
- False negative (FN) – person having disease and the prediction also has a positive
- TP and TN can be used to calculate accuracy rate and the error rates can be computed using FP and FN values.
- True positive rate can be calculated as TP by a total number of persons have disease in reality.
- False positive rate can be calculated as FP by a total number of persons do not have disease in reality.
- Precision is TP/ total number of person have prediction result is yes.
- Accuracy is the total number of correctly classified records.

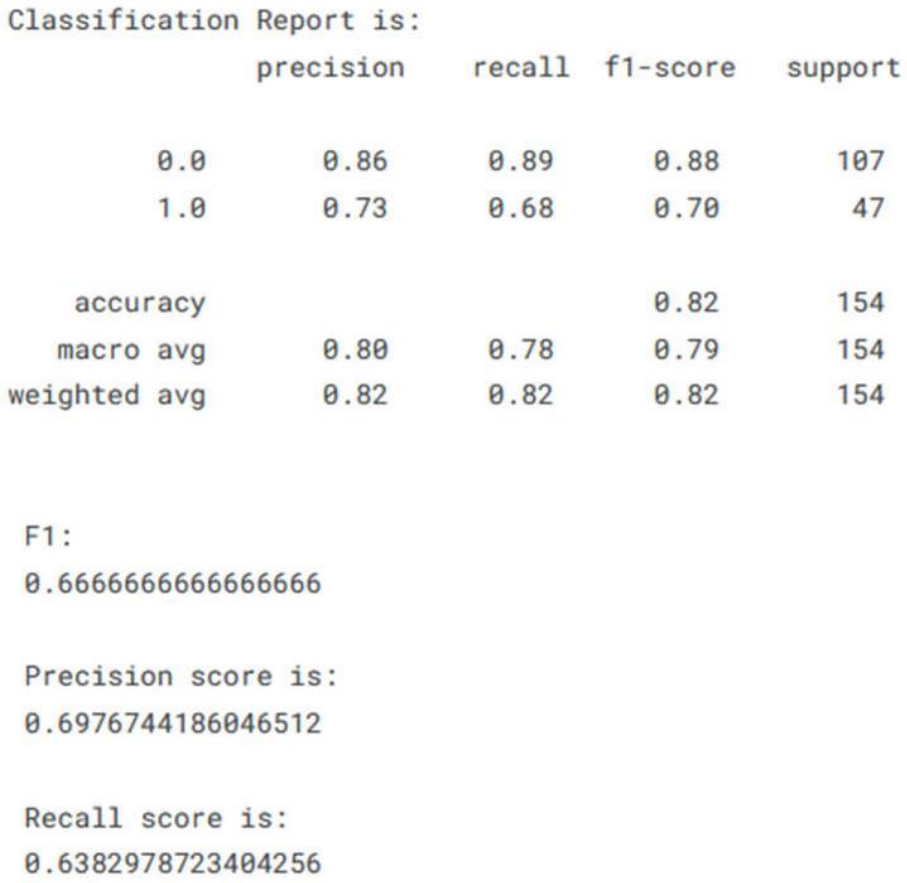
Accuracy- We have chooses accuracy matrix to measure the performance of all the models. The ratio of number of correct predictions to the total number of predictions Made.

$$\textbf{Accuracy} = \frac{\textit{Number of correct Prediction}}{\textit{Total numbers of predictions made.}}$$

6. Results and Analysis

The project predicts the onset of diabetes in a person based on the relevant medical data is collected. When the person enters all the relevant medical data required in the online Web portal, this data is then passed on to the trained model for it to make predictions whether the person is diabetic or non-diabetic the model then makes the prediction with an accuracy of 98%, which is fairly good and reliable. Following figure shows the basic UI form which requires the user to enter the specific medical data fields. These parameters help determine if the person is prone to develop diabetes our research has the added benefit of an associated Web app, which makes the model more user friendly and easily understandable for a novice.

Basic Design of UI



```
Classification Report is:
              precision    recall  f1-score   support

         0.0       0.86      0.89      0.88       107
         1.0       0.73      0.68      0.70        47

    accuracy                           0.82       154
   macro avg       0.80      0.78      0.79       154
weighted avg       0.82      0.82      0.82       154


 F1:
 0.6666666666666666

 Precision score is:
 0.6976744186046512

 Recall score is:
 0.6382978723404256
```

7. CONCLUSION

The objective of the project was to develop a model which could identify patients with diabetes who are at high risk of hospital admission. Prediction of risk of hospital admission is a fairly complex task. Many factors influence this process and the outcome. There is presently a serious need for methods that can increase healthcare institution's understanding of what is important in predicting the hospital admission risk. This project is a small contribution to the present existing methods of diabetes detection by proposing a system that can be used as an assistive tool in identifying the patients at greater risk of being diabetic. This project achieves this by analysing many key factors like the patient's blood glucose level, body mass index, etc., using various machine learning models and through retrospective analysis of patients' medical records. The project predicts the onset of diabetes in a person based on the relevant medical details that are collected using a Web application. When the user enters all the relevant medical data required in the online Web application, this data is then passed on to the trained model for it to make predictions whether the person is diabetic or non-diabetic. The model is developed using artificial neural network consists of total of six dense layers. Each of these layers is responsible for the efficient working of the model. The model makes the prediction with an accuracy of 98%, which is fairly good and reliable.

8. REFERENCES

- https://www.kaggle.com/code/shrutimechlearn/step-by-step-diabetes-classification-knn-detailed
- https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a
- https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm
- https://reachmd.com/news/scientists-develop-12-hour-method-to-predict-diabetes-onset-in-patients-using-artificial-intelligence/2447709/