International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 10 | Oct 2020 www.irjet.net p-ISSN: 2395-0072

Use of Machine Learning in Predicting Diabetes

Krishna Mridha[1]

[1](B.th(CE, Marwadi University, Rajkot, Gujarat, India))

Abstract - Diabetes mellitus, commonly known as diabetes, is a metabolic disease that causes high blood sugar. We all know that the hormone normally passing through the blood to our cells for gathered energy. Any Diabetes contains unprocessed high blood that can damage our nerves, eyes, kidneys, hearts, and other organs. In certain times, there are a lot of people suffer or affected by diabetes for the reason the people have to go to the diagnostic center, hospital, or clinic for tests. So naturally, the management has to store the required tests report and provide a proper diagnosis based on them report. But the rise in machine learning approaches solves this critical problem. Ml(Machine Learning), DS(Data Science) and AI (Artificial Intelligence) play a momentous role in hospitals, diagnostic, clinic, or any Healthcare industries. This type of place must behave a lot of large volume databases. Using a Data analysis technique, one can study a huge dataset and find deeper information, deeper patterns, deeper symptoms from the data, and predict outcome accordingly. The intension of this recitation is to scheme a unique model that can predict diabetes with maximum accuracy. In the existing method, the classification and prediction accuracy is not that high. In this paper, I have proposed a new diabetes prediction model that model I already deployed on the server. For getting the best diabetes prediction I include some extra inputs along with regular inputs such as Age, Glucose, BMI, etc. There are six classification of ML (machine learning) algorithms such as Gradient Boosting Classifier, Logistic Regression, Decision Tree, SVM, K-Nearest neighbors, and Naive Bayes are used in this experiment to detect diabetes at an early stage. n this case study I used Pima Indian Diabetes Database (PIDD) which is collected from the UCI machine learning repository. The accuracy of all the six algorithms is measured on different techniques like Precision, Accuracy, F-Measure, and Recall. From this study, I got the highest accuracy of 83.76% from Gradient Booster Classification comparatively than any other algorithms using in this experiment. These Accuracies are corroborated apply Receiver ROC (Operating Characteristic Curve) in an appropriate way.

Key Words: Data Analysis, Artificial Intelligence , Classification, PIDD, Healthcare

1. INTRODUCTION

Diabetes is nothing but a situation that spoil the ability of body's to process blood glucose, otherwise known as blood sugar. Now in India Over 30 million people have been affected by diabetes. In the Indian urban area the Crude prevalence rate (CPR) is to be 9% but in rural area is approximately 3% of total population where Indian total population more than 1000 million. The implications for the Indian healthcare system are enormous. Type I diabetes: Also known as juvenile diabetes, this type occurs when the body fails to produce insulin. In this study i want to mention three types of diabetes based on (PIDD). Type 1 : Types 1 diabetes generally are insulin dependent that means these affected must be take up artificial insulin daily to live on. Type -2: The lifelong diseases is called type 2 diabetes. On the other hand which people are affected type-2 diabetes they are mentioned to have resistance of insulin. Gestational diabetes does not occur in all women and usually resolves after giving birth.

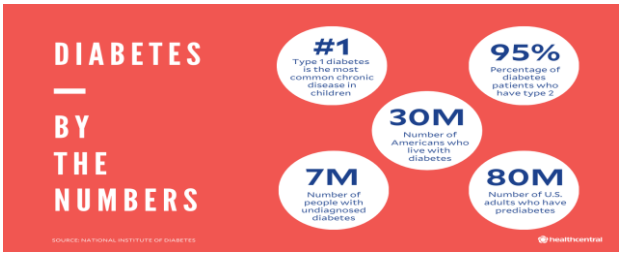


Figure 1: Diabetes by the numbers

2. OVERVIEW OF MACHINE LEARNING

Machine learning is nothing but a utilization of Artificial Intelligence (AI). Machine Learning is a process of learning automatically and amend from last experiences with being implicitly programmed. ML actually direction on the evolution of computer programs that can analysis data and apply it acquire forthemselves. Traditionally, software engineering combined human created rules with data to create answers to a problem.

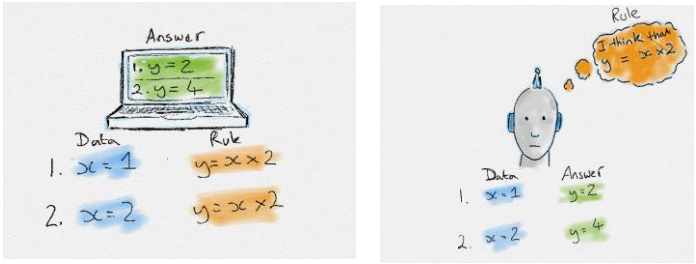


Figure 2: Traditional Programming vs Machine Learning

Instead, machine learning uses data and answers to discover the rules behind a problem. (Chollet, 2017).Most of us are unaware that we already interact with Machine Learning every single day. Every time we Google something, listen to a song or even take a photo, Machine Learning is becoming part of the engine behind it, constantly learning and improving from every interaction. It's also behind world-

changing advances like as diabetes prediction. The remaining of the research discussion is organized as follows: Section- 3 briefs Related Work of various classification techniques for prediction of diabetes, Section-4 describes the Methodology and brief discussion of Dataset used, Section-5 evaluated, Section-6 Results, and Section-7 determines the Conclusion of the research work.

3. RELATED WORK

Dr Saravana Kumar N M, Eswari, Sampath P and Lavanya S (2015) implemented a system using Hadoop and Map Reduce technique for analysis of Diabetic data. This system predicts type of diabetes and also risks associated with it. The system is Hadoop based and is economical for any healthcare organization.**[4]**

Aiswarya Iyer (2015) used classification technique to study hidden patterns in diabetes dataset. Naïve Bayes and Decision Trees were used in this model. Comparison was made for performance of both algorithms and effectiveness of both algorithms was shown as a result.**[5]**

K. Rajesh and V. Sangeetha (2012) used classification technique. They used C4.5 decision tree algorithm to find hidden patterns from the dataset for classifying efficiently**. [8]**

Humar Kahramanli and Novruz Allahverdi (2008) used Artificial neural network (ANN) in combination with fuzzy logic to predict diabetes.**[9]**

B.M. Patil, R.C. Joshi and Durga Toshniwal (2010) proposed Hybrid Prediction Model which includes Simple K-means clustering algorithm, followed by application of classification algorithm to the result obtained from clustering algorithm. In order to build classifiers C4.5 decision tree algorithm is used.**[10]**

Mani Butwall and Shraddha Kumar (2015) proposed a model using Random Forest Classifier to forecast diabetes behaviour.**[7]**

Nawaz Mohamudally1 and Dost Muhammad (2011) used C4.5 decision tree algorithm, Neural Network, K-means clustering algorithm and Visualization to predict diabetes**. [11]**

4. MODELING AND ANLAYSIS

4.1: Proposed Model

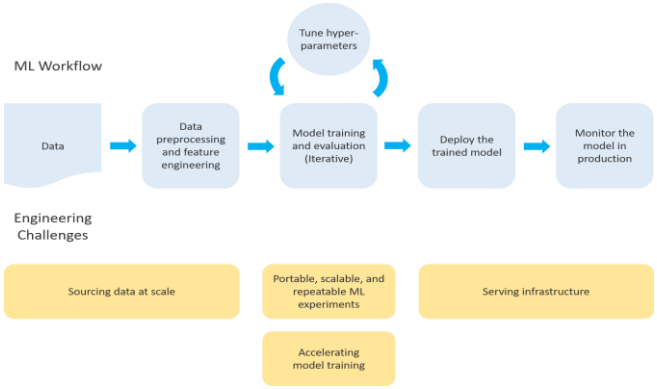


Figure 3: Proposed Model Diagram

4.2: ABOUT USING ALGORITHMS:

In this portion we are going to discuss our top three algorithms step by step along with one image.

4.3: SUPPORT VECTOR MACHINE

Support Vector Machine also known as svm is a supervised machine learning algorithm. Svm is most popular classification technique. Svm creates a hyper plane that separate two classes. It can create a hyper plane or set of hyper plane in high dimensional space. This hyper plane generally utilized for both classification and regression.

a. First we have to select the hyper plane based on the class divides better.

B .For finding the good best hyper plane we have to find the Margin

c. If the distance between the classes is low then the chance of miss conception is high and vice versa.

d. Then select the high margin class. (Margin = distance to positive point + Distance to negative point.)

4.4: GRADIENT BOOSTER CLASSIFIER:

Using the Gradient Booster Classifier algorithm the initial accuracy is attained as 0.764 but after applying feature selection methodology the final accuracy gets increased to 0.84 to a greater value increasing the overall speed of code execution..

a. Consider a sample of target values as P

b. Estimate the error in target values.

c. Update and adjust the weights to reduce error M.

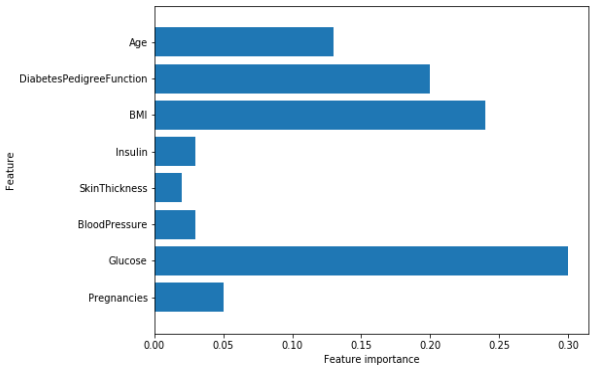


Figure 4: Feature_importance in Gradient Booster

If we see the feature importance graph, the Gradient Booster Classifier gives a lot of importance to the "Glucose" feature, but it also chooses "BMI" to be the 2nd most informative feature overall.

4.5: LOGISTIC REGRESSION:

Logistic regression is also a supervised learning classification algorithm. It is used to estimate the probability of a binary response based on one or more predictors. They can be continuous or discrete. Logistic regression used when we want to classify or distinguish some data items into categories. It classify the data in binary form means only in 0 and 1 which refer case to classify patient that is positive or negative for diabetes. Main aim of logistic regression is to best fit which is responsible for describing the relationship between target and predictor variable. Based On linear regression, logistic regression is created. Logistic regression model uses sigmoid function to predict probability of positive or negative. Sigmoid function P = 1/1+e - (a+bx) Here P = probability, a and b = parameter of Model.

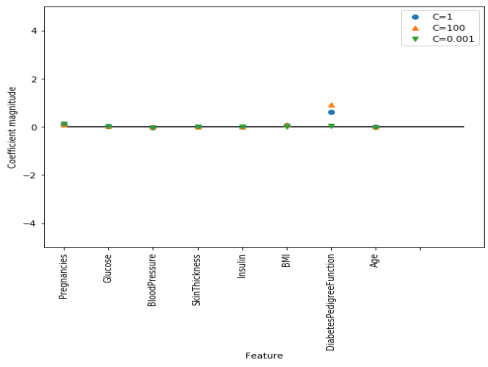


Figure 5: Features

4.6: K-NEAREST NEIGHBORS:

One of the most imoortance algrothm is K-Nearest Neighbors. It 's belongs to to the supervised learning, finds intense application in pattern recognition, data mining and intrusion detection.The K-Nearest Neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.

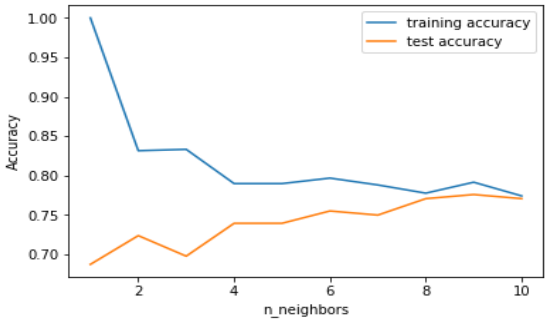


Figure 6: Training and Test accuracy in K-neighbors (Histogram)

The above plot shows the training and test set accuracy on the y-axis against the setting of n_neighbors on the x-axis.

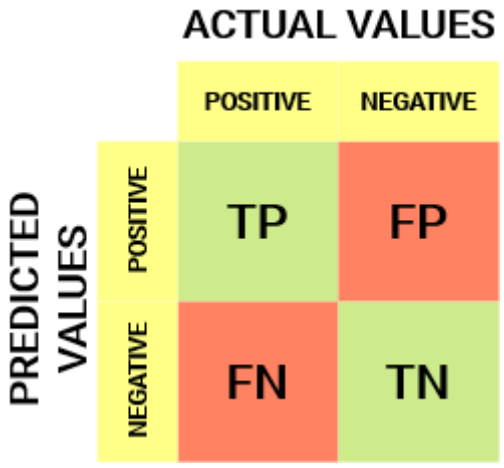
5. EVALUATION

Through the classification accuracy, confusion matrix, f1-measure, precision, and recall, we can predict the results.

Classification Accuracy: Classification is the ratio of number of accurate predictions and the total number of samples input.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

Confusion Matrix: Confusion matrix in nothing but an n * n matrix which is user for calculate the accuracy of any classification model values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making.



Let's see how our model performed:

Table 1

ID	Actual Sick?	Predicted Sick?	Outcome
1	1	1	TP
2	0	0	TN
3	0	0	TN
4	1	1	TP
5	0	0	TN
6	0	0	TN
7	1	0	FP
8	0	1	FN
9	0	0	TN
10	1	0	FP
:	:	:	:
1000	0	0	FN

F1-Score: For testing accuracy we used f1-score. What ever we got recall and precision, f1-score is the harmonic between of them. The range for f1 score is 0,1. Maninly it's used to find out the balance between precision and recall. Mathematically, it is given as-

$$F1 - score = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

F1 Score tries to find the balance between precision and recall.

Precision: Precision tells us how many of the correctly predicted cases actually turned out to be positive.

$$Precision = \frac{TP}{TP + FP}$$

Recall: Recall tells us how many of the really positive cases we were able to predict correctly with our model

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

Data Set

The proposed methodology is evaluated on Diabetes Dataset namely (PIDD), which is taken from UCI Repos-itory. This dataset consist of 768 female patients instance. The dataset also comprises numeric-valued 8 attributes where value of one class '0' treated as tested negative for diabetes and value of another class '1' is treated as tested positive for diabetes.

Attribute Description:

Table 2

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

6. RESULT AND DISCUSSION

The final result has been collected successfully using the machine learning algorithms. There is different accuracy related to different machine learning algorithms used such as Random forest classifier, Support vector machine, Logistic Regression, K-Nearest Neighbour, Gaussian Naïve Bayes, Gradient Boost Classifier. According to the Indian pima diabetes dataset available in UCI Repos-itory. After using different machine learning algorithms of PIDD dataset we collectected accuracies as mentioned below. Gradient Booster Classification algorithm provide best or top performance or accuracy of 84%.

Table 3. Accuracy Table

Algorithms	Accuracy
Gradient Booster Classifier	84%
Support Vector Machine	81%
Logistic Regression	83%
K-Nearest Neighbour	80%
Gaussian Naïve Bayes	79%
Random Forest	80%

Confusion Matrix Heat map for Gradient Booster Classification is given below-

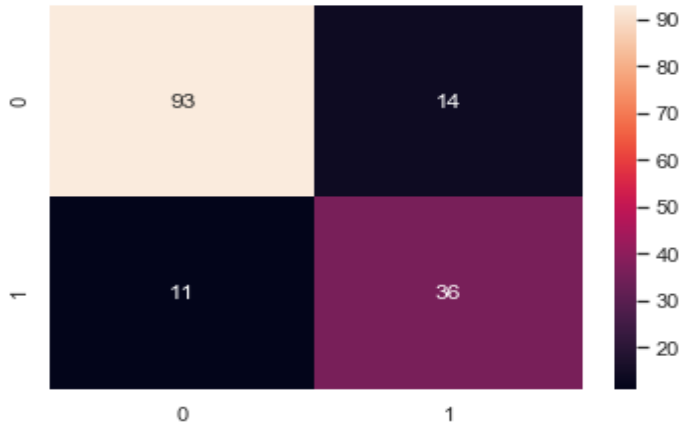
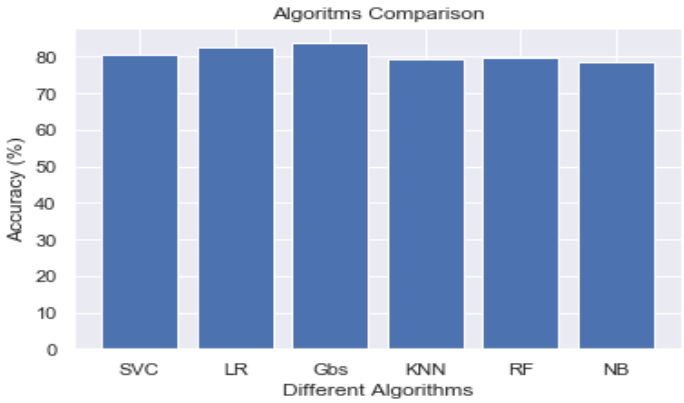


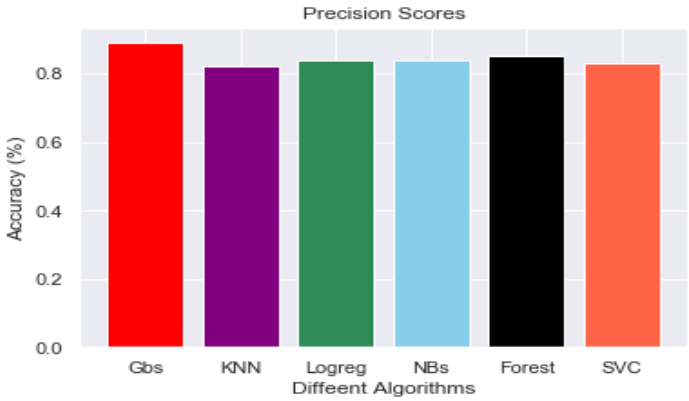
Figure 7: Heatmap of Confusion Matrix

If you follow the Table 3, you can easy find out the best accuracy with the name of classification algorithm.



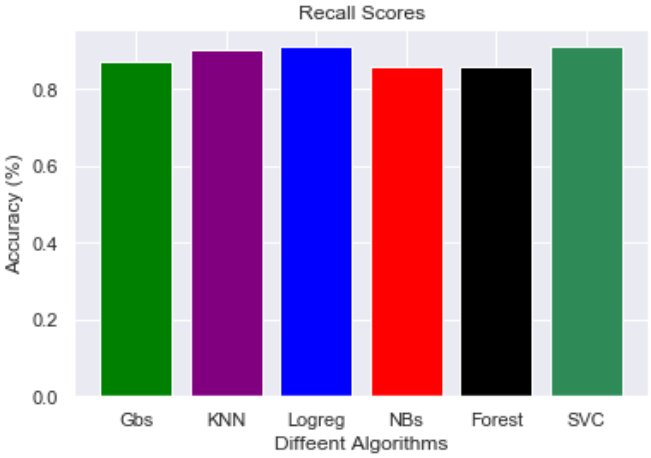
	KNN	SVC	Gradient B.C	Logistic	Naïve Bays	Random F.
ROC	80%	81%	84%	83%	79%	80%

Figure 8: Accuracy Comparison.



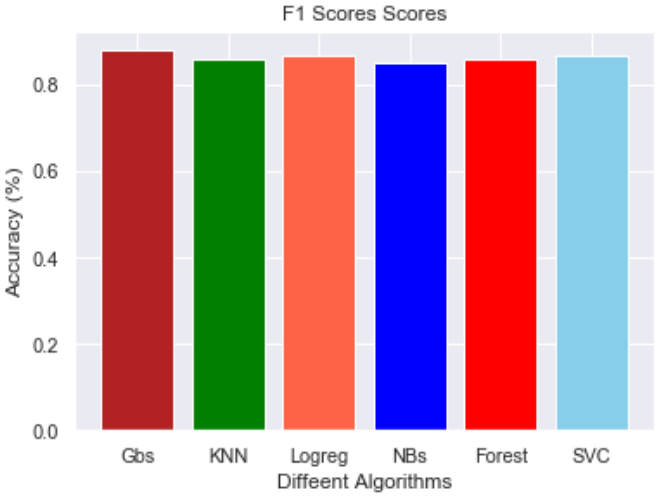
	Gradient B.C	KNN	Logistic	Naïve Bays	Random F.	SVC
Precision	0.89	0.82	0.84	0.84	0.85	0.83

Figure 9: Precision



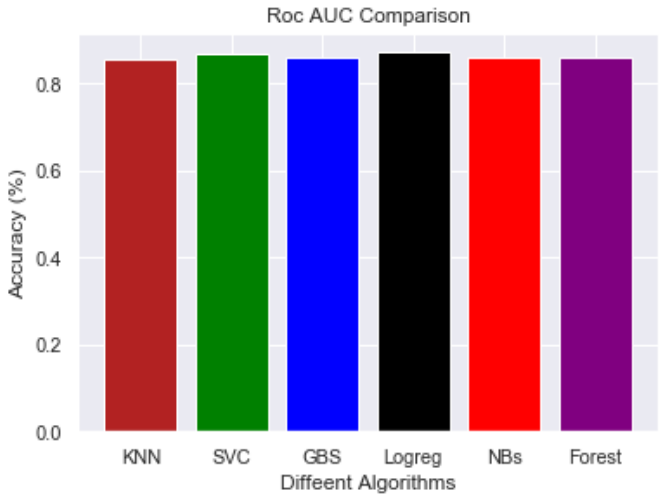
	Gradient B.C	KNN	Logistic	Naïve Bays	Random F.	SVC
Recall	0.87	0.90	0.91	0.86	0.86	0.91

Figure 10: Recall



	Gradient B.C	KNN	Logistic	Naïve Bays	Random F.	SVC
F1_measures	0.88	0.86	0.87	0.85	0.86	0.87

Figure10:F1_measures



	KNN	SVC	Gradient B.C	Logistic	Naïve Bays	Random F.
ROC	0.86	0.87	0.87	0.87	0.86	0.86

Figure 11: ROC area

Deployment the model on Server (Web Application)

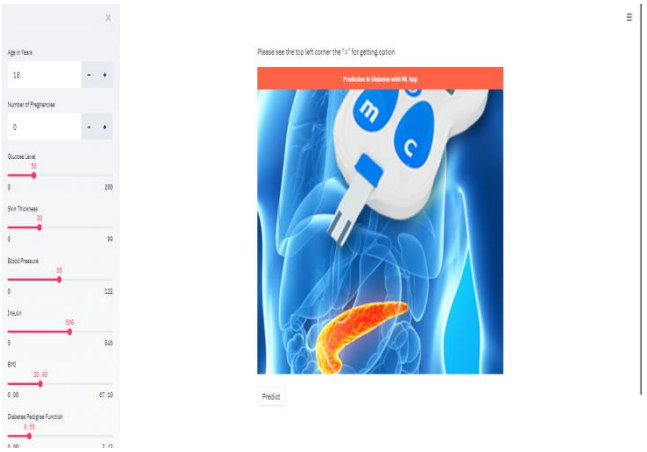
Deployment the model on Server (Web Application)

 Input: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Age,

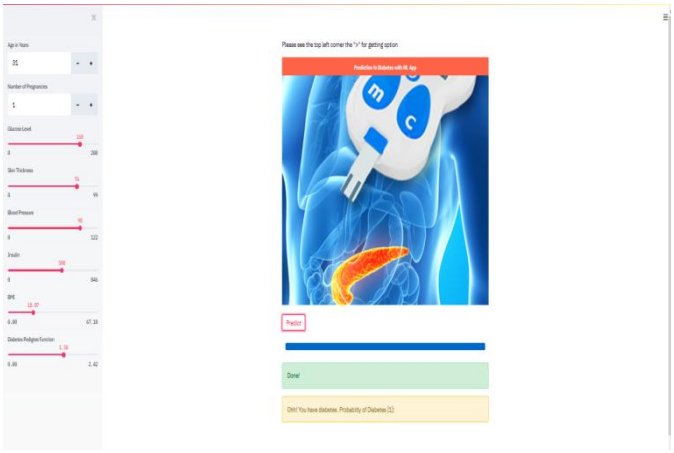
 Output: Predict the diabetes

- When this concept is applied development, we can create a web application that's imply takes 8 inputs from the user to result of diabetes.
- Here is how a web application created using python, Streamlit looks like:

Scenario 1: We can able to see a arrow "<" sign top left side When input are Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin



Scenario 2: After Inputting 8 values, we have to press "predict" button and have to wait to complete progress bar as well as the output.



App Url Link: https://diabetes-streamlit-app.herokuapp.com/

Souce Code Uril Link: https://github.com/krshna-mridha/Diabetes_Streamlit_App

7. CONCLUSION

At last we have reached our results of predicting diabetes by applying machine learning algorithms widely. The entire procedure is nothing but a machine learning process.. We have completed a classification model task with web or app development it can turn in into application of a large detect wildfires all around the globe.

ACKNOWLEDGEMENT

It gives us great pleasure in presenting the preliminary project report on '**Use of Machine Learning in Predicting Diabetes**'. One of my university professor help me a lot in sometimes. And my roommates are helping to chose color and design even image for UI design.

8. REFERENCES

[1] Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference On I-SMAC,978-1-5090-3243-3,2017.

[2] Ayush Anand and Divya Shakti," Prediction of Diabetes Based on Personal Lifestyle Indicators", 1st International Conference on Next Generation Computing Technologies, 978-1-4673-6809-4, September 2015.

[3] B. Nithya and Dr. V. Ilango," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7,2017.

[4] Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing,2015.

[5] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly," Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.

[6] P. Suresh Kumar and S. Pranavi "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics", International Conference on Infocom Technologies and Unmanned Systems, 978-1-5386-0514-1, Dec. 18-20, 2017.

[7] Mani Butwall and Shraddha Kumar," A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier", International Journal of Computer Applications, Volume 120 - Number 8,2015.

[8] K. Rajesh and V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.

[9] Humar Kahramanli and Novruz Allahverdi,"Design of a Hybrid System for the Diabetes and Heart Disease", Expert Systems with Applications: An International Journal, Volume 35 Issue 1-2, July, 2008.

[10] B.M. Patil, R.C. Joshi and Durga Toshniwal,"Association Rule for Classification of Type-2 Diabetic Patients", ICMLC '10 Proceedings of the 2010 Second International Conference on Machine Learning and Computing, February 09 - 11, 2010.

[11] Dost Muhammad Khan1, Nawaz Mohamudally2, "An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm", Journal Of Computing, Volume 3, Issue 12, December

BIOGRAPHIES

Krishna Mridha is a Computer Engineering student of Marwadi Univesity in Gujarat. He is very interested to learn Machine Learning and Deep Learning.