International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 11, Issue 3 [May.-June. 2022] PP: 82-87

Real Time Diabetes Prediction

P.Bharathi[1], M.Pankaj Sai[2], M.S.S.Ramakrishna[3], K.Sravan Kumar[4], H.Shirisha[5], G.Manoj[6]

[1]*Assistant Professor, Department of CSE, Raghu Institute of Technology, Vizag,India*
[2,3,4,5,6]*Student, Department of CSE, Raghu Institute of Technology, Vizag, India*

***ABSTRACT**: "Real-time diabetes prediction" is a web application that helps in the prediction of diabetes. Diabetes is an enduring sickness that can cause an overall well-being disaster. The International Diabetes Federation Council (IDF) has proven to us that there were three thousand eight hundred and twenty lakhs people in the world are living with diabetes and this number can be twice in the next 15 years. Diabetes, otherwise called Diabetes Mellitus is an ongoing sickness caused because of the increment of glucose levels in the blood. If diabetes is ignored this raises severe health issues in the human body such as kidneys, heart, eyes, foot, nerves, and even death. we can avoid this from happening if we detect diabetes early.*

***Keywords:** Machine learning, Diabetes, XGBoost, Pima Indian dataset, Diabetes.*

Date of Submission: 25-05-2022 Date of Acceptance: 05-06-2022

I. INTRODUCTION

The World Health Organization (WHO) reported that around 16 lakhs pass due to diabetes every year. Diabetes is a kind of disease that occurs when the blood glucose/blood sugar level in the human body is remarkably high. According to doctors, diabetes ensues when the human body's gland called the pancreas cannot make sufficient insulin (Type 1 diabetes) for this case of diabetes there is no permanent cure. The produced insulin cannot be used by the body (Type 2 diabetes) this is caused by both living matters and heredity. There is another type of Diabetes which is gestation diabetes which is when women get pregnant. After processing of food that we ate, glucose gets released. Insulin is a hormone that transfers from one blood cell to another cell and guides to use of blood glucose and convert that into energy. If the pancreas fails to produce sufficient insulin, the body cannot convert glucose into energy and that glucose remains in the blood. Thus, the blood glucose/blood sugar raises in the body at a very high level. If a person is suffering from high blood sugar, some signs(symptoms) are shown in the human body like excessive hunger, frequent urination, and intense thirst. The normal level of glucose for a healthy person ranges from 70 milligrams per deciliter to 99 milligrams per deciliter. For abnormal persons, this is greater than 126 milligrams per deciliter, which indicates diabetes. Prediabetes is a stage when a person has a glucose level between 100 and milligrams per deciliter (mg/dl). if glucose level raises this may cause health problems like kidney failure, heart disease, stroke, amputation, eye, and nerve damage, Tired or sleepiness, losing weight, mood swings, and Regular infections. Till today there is no permanent cure for diabetes. Predicting an individual's risk and vulnerability to a chronic illness like diabetes is an important task. The efficient control of diabetes is possible if it can be detected early. Our project helps in achieving this using machine learning algorithms.

II. LITERATURE SURVEY

There are other scholars who used the ML (machine learning) models to predict diabetes using Pima Indian diabetes records. The PIDD (Pima Indian Diabetes dataset) has 9 variables, 768 tuples describing female patients. Some similar and previous works are discussed in this section.

J. Beschi Raja et.al [1] in this work use different types of classification on the dataset. These models are done to tell whether a person may get diabetes or not in the future by given attributes. This group collected information from the hospital warehouse. Some information is gathered by doctors of that hospital this information came helpful for them as well as for us. In this, they used Weka to pre-process the data they used different methods like the random forest, and Gradient boost In which Gradient boost performed well with an accuracy of 89.7%.

K.Pavani et.al [2] aim to measure the classification not only based on accuracy but also with recall, precision level, and F-measure. In this, they used many models for prediction. Seven models were used to be precise. These models are logistic regression, support vector machine, random forest, decision tree, KNN, Naïve Bayes, and gradient boot among which they got the best results in random forest. They divided the data set into

a 75:25 split. They got the best accuracy in both random forest and naïve Bayes with 80% in which they concluded the best accuracy is random forest based on the precision level, recall, and f measure.

Jobeda Jamal Khanam et. al [3] have applied a soft-computing technique for an intelligent diagnosis of diabetes prediction. They used weka for pre-processing all models they used scored accuracy greater than 70% among which they got the best accuracy with ANN with 88.6% they have proven that with a better pre-processing same model can achieve the better result they got better results than previous models for the same classification. This research hopes to get better accuracy than previous models which will give hand in the health industry for better cures.

Niloy sikder et. al [4] this project use different approaches using ensemble learning which is stacking, boosting, and bagging in this they crossed 90% up to know this is the best model that we saw. Through this, we got an idea of using different ensembles to increase accuracy. In this model, they got an accuracy of 92%. This model gave us an idea of which model we should be using in this project.

Table 1 Comparison with previous work

Year	Work	Approach	Accuracy
2019	J. Beschi Raja et.al [1]	Randomforest, Gradinent boosting	(82.2%), (89.7%)
2020	K.Pavani et.al[2]	Random forest, Naïve bayes	(80.0%), (79.8%)
2021	Jobeda Jmal Khanam et.al[3]	Artificial Neural Network	88.6%
2021	Niloy sikder et. al[4]	Ensemble Learning	92%
2022	This project	XGB, LightGBM	(95.4%),(91.5)

III. METHOD AND SYSTEM ARCHITECTURE

In this project, we are using python and its powerful modules/libraries to predict diabetes. Starting from importing datasets to applying various algorithms are done in python for this we are using the most powerful and the best idle from google which is google colab. It is a cloud based environment it uses the most powerful CPUs and GPUs at remote locations and gives us fast and accurate results. It is easy to code in colab as it supports popular machine learning modules.

3.1 Machine learning architecture

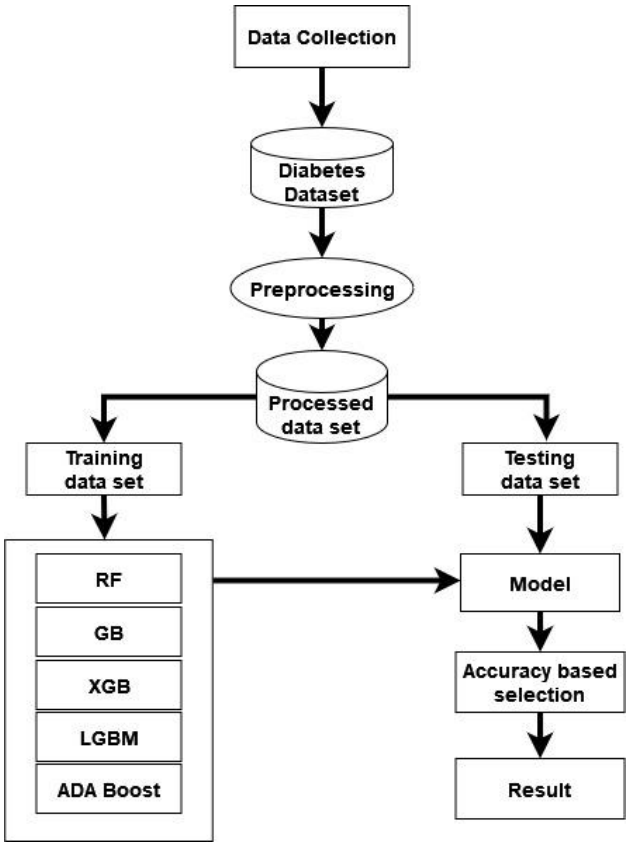

Figure 1 Machine learning model architecture

3.2 System architecture

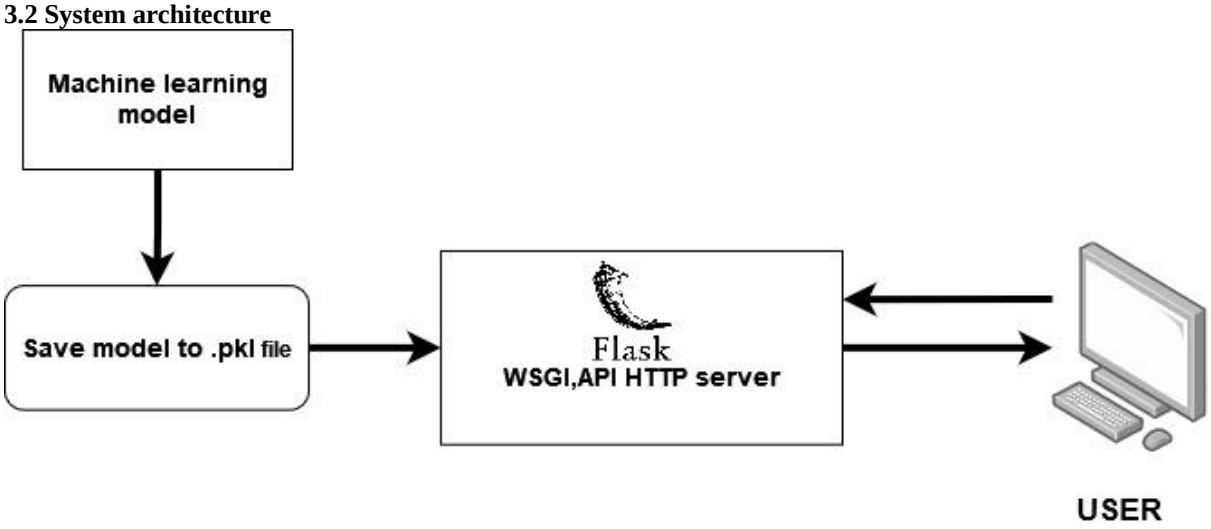

Figure 2 System architecture

IV. Data Pre-processing

4.1 Data collection and description

For this project, we used a dataset that is made available to the public by the National Institute of Diabetes and Digestive and Kidney Diseases. The name of the dataset is PID (Pima Indian dataset). Dataset plays a very important role in machine learning. Collection of the data is very hard as we need to do many surveys, check records, and many other sources and modes of information. Once we acquire the dataset that we need, the first thing we need to do is data pre-processing. The data which we took have 768 tuples and 9 variables. Those attributes are:

a) Pregnancies
b) Glucose
c) Blood Pressure
d) Skin Thickness
e) Insulin
f) BMI
g) Diabetes Pedigree Function
h) Age
i) Outcome

Table 2 Description Data set

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

From the above table, We can observe that in Pregnancies column, it shows us how many times a woman got pregnant. Glucose shows us Plasma glucose concentration for 2 hours in an oral glucose tolerance test. Even in the blood pressure column, it shows us that the Diastolic blood pressure is millimeters of mercury.

Skin thickness shows us triceps skinfold thickness in millimeters. BMI shows us the body max index. Insulin shows us 2-Hour serum insulin in a micro international unit of substance per milliliter. Ages show usage of that person.

4.3 Identifying and handling missing data

In the above mention data, there are only a few zero values. If there are zero values in our data this may affect our prediction. So, we need to make this inconsistent data to consistent for that we use a few data preprocessing techniques. The below graph shows us the count of zeros in each attribute. At first, we need to convert our zero value to Null so that the data preprocessing becomes easy. Data pre-processing is important to get accurate results. Finally, we need to fill in the null values with mean, median, mode, or the standard deviation if the tuple has more than 25% of the data. We remove the entire tuple if 75% of the data is not present.

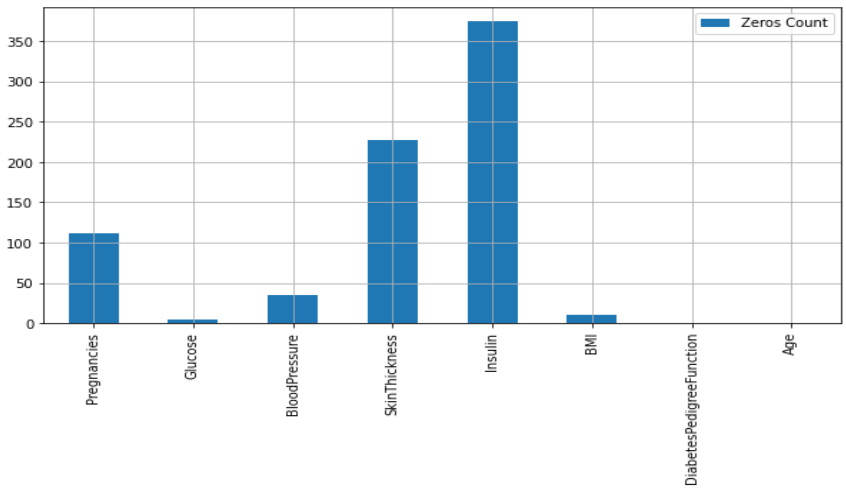

Figure 3 Missing data bar graph

In the above graph, we can see the count of zero values in each attribute. We need to identify the dependent variable with the independent variable. Now we replace data with a median. In our research, we found that the median is the best fit for this dataset and after replacing our null data with the median we need to split our data

4.4 Splitting of data

After handling our noisy or inconsistent data we split our data into a training set and a testing set. Most of the time we split our data into either 70:30 or 80:20 ratios. For this project, we only divided our dataset into an 80:20 ratio and that means eighty percent of the dataset is training data and the remaining twenty percent of our data is testing the data.

4.5 Applying ML Algorithm

In this paper, we have used over five machine learning algorithms. i.e Random Forest (RF), Gradient Boost (GB), eXtreme Gradient Boosting (XGBoost), Light Gradient Boost model(LGBM), Adaptive Boosting (ADA). After performing these ML algorithms on our dataset, we noted the outcomes of each classifier. And we observed that Extra Tree Classifier outperforms when compared other machine learning models that we have used in every aspect such as Accuracy, Recall, Precision, and F1-Score.

V. Result and analysis

After applying the above models to the dataset we got our accuracy results. Through this, we determine which model will be best suited for our project. From the table, we see that the extreme gradient boost algorithm performs well than the remaining machine learning model.

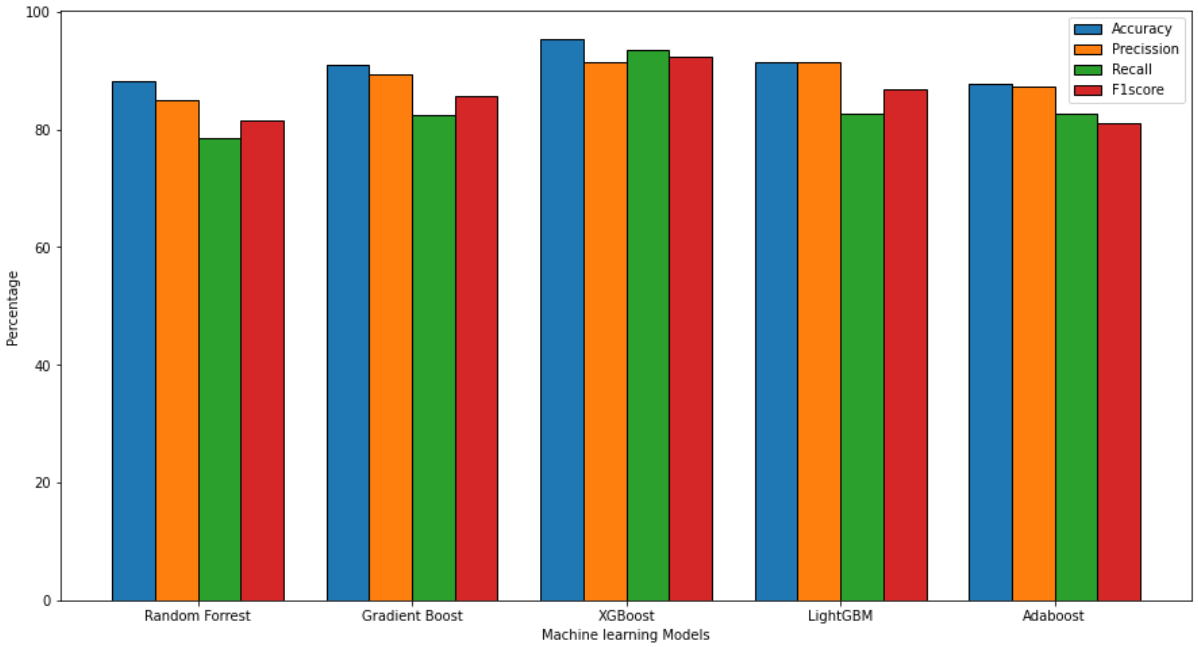

Figure 4 Result Bar graph

Table 3 Result Table

Models	Accuracy	Precision	Recall	F1-Score
Random Forest	88.3%	87.2%	77.3%	82.0%
Gradient Boost	90.2%	89.3%	80.7%	84.8%
XG Boost	95.4%	91.4%	93.4%	92.4%
Light GBM	91.5%	91.4%	82.6%	86.8%
Ada Boost	87.6%	87.2%	75.9%	81.1%

5.1 Home Page

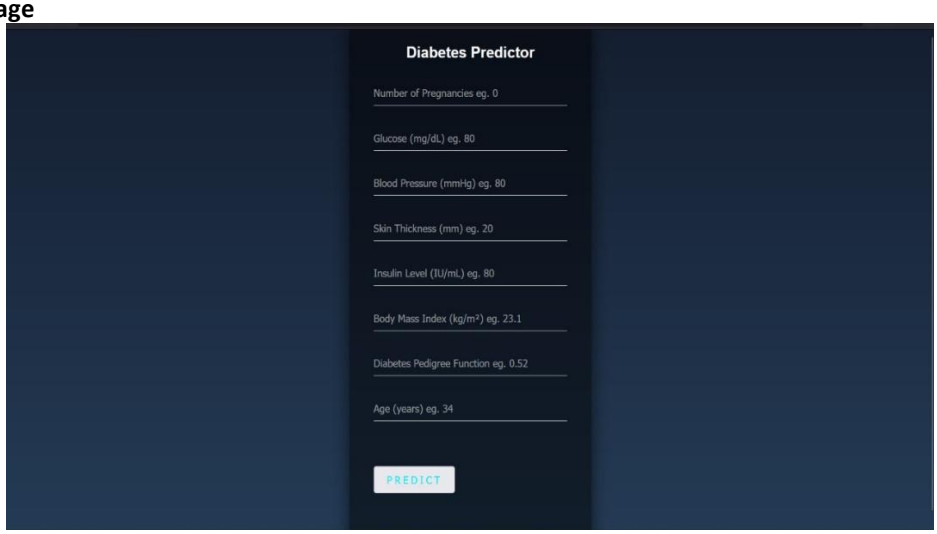

Figure 5 Home page

5.2 Result Page

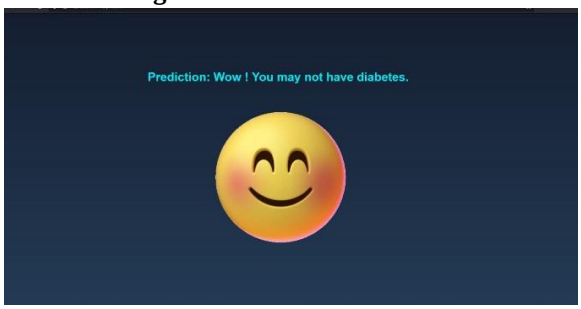

Figure 6 Result page 1

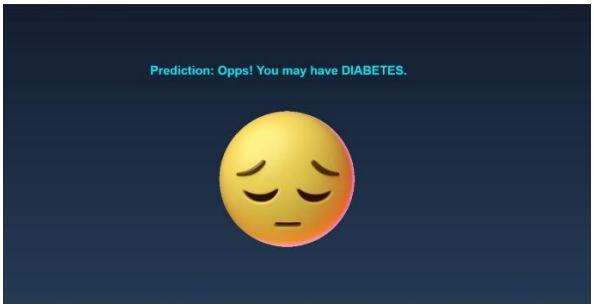

Figure 7 Result page 2

After taking inputs from the Homepage (Figure 5) we will get Result page 1 (Figure 6) if the particular patient does not have diabetes else we get the Result page 2 (Figure 7) if the particular patient is diabetic.

VI. Conclusion

The principal part of this work is to make an effective diagnosis system for diabetes prediction using data set of diabetes patients utilizing five distinctive ensemble machine learning classifiers. We researched all classifier's execution of patients' information parameters and the XgBoost classifier gives us an accuracy of 95.4% not only accuracy but also outperforms other measurements like precision, recall, and f1socore this is the highest accuracy that we achieved. If we have a proper dataset for another disease we can predict those diseases using machine learning classification in a similar way.

Reference

[1]. J. Beschi Raja et.al https://www.ijeat.org/portfolio-item/A9898109119/
[2]. K.Pavani et.al https://link.springer.com/chapter/10.1007/978-981-15-5089-8_41
[3]. Jobeda Jmal Khanam et.al https://www.sciencedirect.com/science/article/pii/ S2405959521000205
[4]. Niloy sikder et.al https://www.mdpi.com/2073-8994/13/4/670
[5]. WHO (world health organization) https://www.who.int/health-topics/diabetes
[6]. Dataset of Pima Indian dataset from Kaggle repository
[7]. https://www.medicalnewstoday.com/articles/325018#how-is-the-pancre as-linked-with- diabetes.
[8]. Meigs JB, D'Agostino RB Sr, Wilson PW, Cupples LA, Nathan DM, Singer DE. Risk variable clustering in the insulin
[9]. Anna V, van der Ploeg HP, Cheung NW, Huxley RR, Bauman AE. Sociodemographic correlates of the increasing trend
[10]. Bellamy L, Casas JP, Hingorani AD, Williams D. Type 2 diabetes mellitus after gestational diabetes: a systematic review
[11]. https://www.webmd.com/diabetes/diabetes-causes.
[12]. https://www.mayoclinic.org/diseases-conditions/prediabetes/ diagnosis-treatment/drc-20355284
[13]. https://codepen.io/soufiane-khalfaoui-hassani/pen/LYpPWda for frontend Html and CSS