IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DIABETES PREDICTION USING MACHINE LEARNING TECHNIQUES

V P Punya[1], Sindhu M N[2], Poornima M E[3], Brindashree[4], Dr. Rajeshwari J[5]

1,2,3,4 Students, Department of Information Science and Engineering, Dayananda Sagar College of Engineering Bangalore, Karnataka, India

5 Associate Professor, Department of Information Science and Engineering, Dayananda Sagar College of Engineering Bangalore, Karnataka, India

Abstract: Diabetes has grown to be a severe problem now a days. So, we need to take severe precautions to eliminate this. To eradicate, we have to predict the occurrence of Diabetes. We predict the occurrence of diabetes with the usage of Random Forest which is Machine Learning Algorithm. Pima dataset is a data of real patients. Using these records, we are able to build accurate model that can predict the existence of diabetes. The factors which we are considering are pregnancies, glucose, blood pressure, skin thickness, insulin, body mass index, diabetes pedigree and age.

Index Terms **– Random Forest, PIMA dataset, Machine Learning.**

I. INTRODUCTION

Diabetes is a disease in which a person will suffer from extended level of blood sugar in their body, due to the deficiency of insulin or if cells in body no longer respond to insulin. The signs and symptoms are numerous organs failure, particularly heart, kidneys, eyes, nerves and veins. The goal of this research is to get the closest outcome. We have proposed a diabetes prediction model for more accuracy using regular factors like pregnancies, blood pressure, glucose, skin thickness, insulin, body mass index, diabetes pedigree, body mass index, age. With new dataset classification the accuracy is boosted when compared to existing dataset. In additional to the diabetes prediction intended in enhancing the accuracy.

II. LITERATURE SURVEY

[1] Author- Sofia Benbelkacem and Baghdad Atmani

Topic – Random Forest for Diabetes Diagnosis

Random forest is one of the latest success studies in finding for Decision tree. It is broadly used in the scientific filed, in particular for diabetes analysis. Diabetes is achieving epidemic proportions so much in developing and newly industrialized countries. Thus, random forest must be exploited to cope with diabetes evaluation. In this paper, we make most of the principle of random forests for the implementation of a powerful version for the analysis of diabetes. The experiments have been achieved on the existing dataset from the Pima Indians dataset is determined from the UCI repository. Then, random forest has been in comparison with different system gaining knowledge of techniques.

[2] Author- V. Anuja Kumari, R. Chitra

Topic- Classification of Diabetes Disease using support vector machine.

In this paper, they have used data sets for diabetes disorder from the ML laboratory at university of California, all of the patients' data are instructed via using SVM. The preference of quality rate of the amount of data given to approach SVM can be efficaciously used to come upon a common region disorder with easy scientific measurements, without test from labs. The overall general performance parameters along with sensitivity, accuracy and specificity of the SVM and making it an outstanding preference for selection.

[3] Author - MingqiLi, XiaoyangFu and Dongdong Li

Topic - Diabetes Prediction Based on XGBoost Algorithm

Pre-processing the data is a vital prerequisite for model accuracy and then XGBoost modelling procedure is made great. The contrast of the algorithms makes it clear that XGBoost is greater green amongst a few conventional algorithms. Through the comparative evaluation with the included set of rules, we proposed the stepped forward function mixture set of rules primarily based totally XGBoost.

[4] Author- Krati Saxena, Dr. Zubina Khan, Shefali Singh.

Topic- Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm.

This paper is about KNN a way that is used for classifying devices based mostly on closest training examples withinside the characteristic space. It assumes to be in n-dimensional space in all instances are elements. A distance degree is needed to determine

the "closeness" of instances. It classifies an example with the useful resources in finding the nearest neighbors and it is also a way that is used for classifying them based on primary and completely on closet training cases within the characteristic space. The most fundamental type of instance based completely learning or lazy learning is KNN.

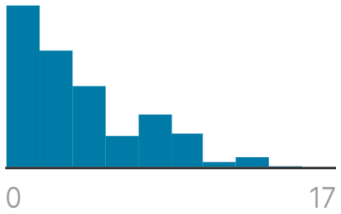
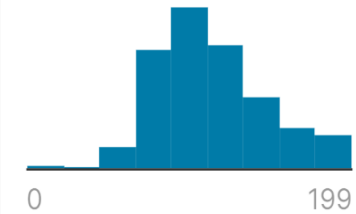
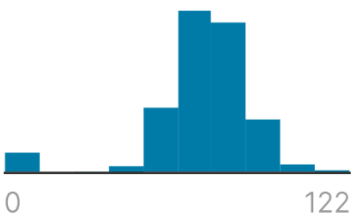
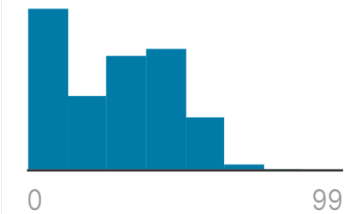
[5] Author- J. Beschi Raja, R. Sujatha, S. Sam Peter, V. Roopa, R. Anitha.
Topic – Diabetes Prediction using Gradient Boosted Classifier
Diabetes is one of the frequent disorders occurred in child and adults. The techniques of Machine Learning enable to perceive the disorder in advance degree for saving it. It is in comparison with ML to know algorithms like Random Forest and Neural Networks. Dataset is hired from Pima Indian Dataset. Then the creation of models is done and they are evaluated via way of means of popular measures which includes AUC, Recall and Accuracy.

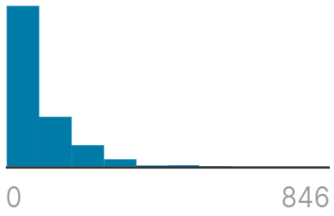
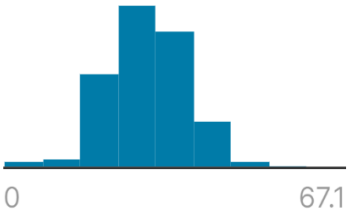
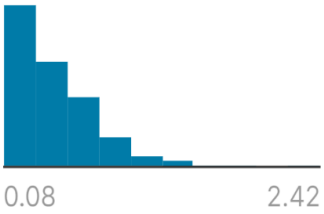
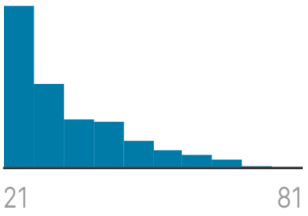
[6] Author- Paratoo RAHIMLOO, Ahmad JAFARIAN.
Topic – Prediction of diabetes by using logistic regression statistical model and artificial neural network and combination of them.
In this paper, attempted via way of means of combining the models which are neural network and statistical and then create a very new compound that has at the least mistakes and most reliability and is analysed. With the above recommendations model and extraordinary studies and comparing, numerical outcomes obtained, the accuracy and performance of the approach has been investigated and acceptable outcomes in comparison to the neural community and logistic regression techniques have been obtained.

III. IMPLEMENTATION

Dataset:
The data set which is used in this project has been taken from Pima Indians diabetes database which is of National institute of diabetes and digestive and kidney diseases. The most important purpose of the dataset used is to verify with diagnostically to predict if patient has diabetes or not based only on the defined diagnostic measurements which includes in the database. The data of all the patients in this particular data set are women at least 21 years old of pima Indian heritage.
The outcome is the dataset in medical predictor and consists only one target variable. As shown in the figures they are independed to each other:

Based on the first 7 independent column values, we are going to predict our model which is machine learning model and then will predict the value of the last column, that is the outcome. 1 and 0 are the two medical values considered here. In that 1 indicates that the patient is diabetic and 0 indicates person is not diabetic.

Pregnancies Number that specifies previous Gestation Periods	**Glucose** Plasma glucose level in an oral glucose tolerance test (2 hrs) (mg/dl)
	
Blood Pressure Pressure in arthritis when the heart rests - Diastolic blood pressure (mm Hg)	**Skin Thickness** Skin fold thickness of Triceps- right arm(mm)
	
Insulin Insulin level in Blood – 2-Hour serum insulin (mu U/ml)	**BMI** Body mass index- measure based on height and weight (weight in kg/ (height in m)^2)

	
Diabetes pedigree function Ancestral diabetic record of a person	**Age** Age of a person(years)
	
Outcome A Class variable that contains 268 entries as 1's and others as 0	

SYSTEM ARCHITECTURE:

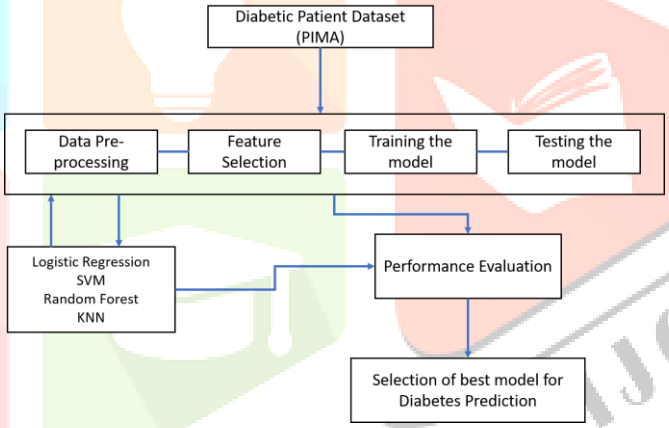


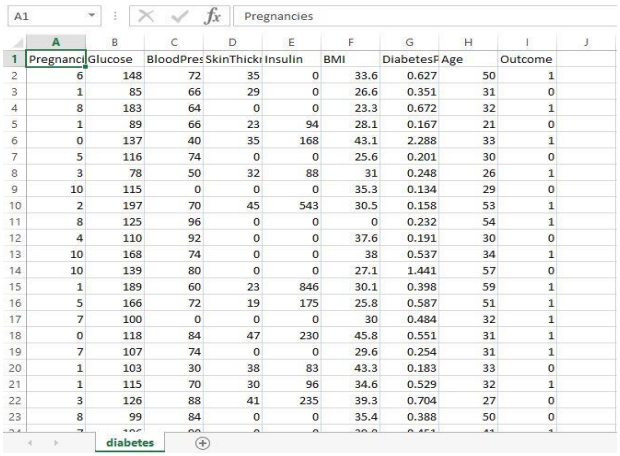
Fig 1: System Architecture

To carry out diabetes prediction, we have taken the existing data set which is from Pima Indian Dataset from Kaggle. This particular data set is taken from the National Institute of Diabetes and Digestive and Kidney Diseases. And then data processing is performed and the data being used will be divided into two sets which are training set and testing set. Now they are sent to ML model which is machine learning model where we have used algorithms like Logistic Regression, KNN, Random forest classifier, SVM. We have selected random forest classifier based on train accuracy and test accuracy which was high compared to all other algorithms which has testing accuracy 93% and training accuracy of 98 %.

Web page creation is done where the user can give user inputs to get the results. The factors to enter in the web page are:

- Pregnancies
- Glucose
- BMI
- Insulin
- Diabetes pedigree
- Blood pressure
- Age
- Skin thickness

Steps in implementation:
1. Training the model using the given dataset.
2. Install the necessary softwares like PyCharm IDE, Django, Anaconda and Web browser.
3. Basic setting on PyCharm
4. Design website home page (Frontend)
5. Design the web page for user input and prediction.
6. Link the trained model to the front end.



	A	B	C	D	E	F	G	H	I
1	Pregnanci	Glucose	BloodPres	SkinThickı	Insulin	BMI	DiabetesF	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10	168	74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100	0	0	0	30	0.484	32	1
18	0	118	84	47	230	45.8	0.551	31	1
19	7	107	74	0	0	29.6	0.254	31	1
20	1	103	30	38	83	43.3	0.183	33	0
21	1	115	70	30	96	34.6	0.529	32	1
22	3	126	88	41	235	39.3	0.704	27	0
23	8	99	84	0	0	35.4	0.388	50	0

Fig 2: PIMA data set

IV. RESULT

Training accuracy: The accuracy of a model based on the dataset it was constructed on.

Algorithm	Accuracy
Logistic Regression	77.03 %
K Nearest Neighbor	82.41 %
SVM	77.19 %
Random Forest	98 %

Test accuracy: The accuracy of a model on the dataset it hasn't seen.

Algorithm	Accuracy
Logistic Regression	82%
K Nearest Neighbor	80.5%
SVM	65.10%
Random Forest	92.18%

V. CONCLUSION

Machine learning techniques were tested and the algorithm with highest accuracy was chosen to predict diabetes. The early detection of diabetes is very important for treatment to lead better and healthy life. Diabetes is the main reason for overweight, physically inactive, stroke, heart attack etc, due to increase in blood glucose level and untreated high blood sugar. Here Diabetes prediction is made with the use of Random Forest Classifier which is very powerful technique

VI. ACKNOWLEDGEMENT

We express sincere thanks to Prof. C P S Prakash, Principal Dayanada Sagar College of Engineering. We would like to extend our gratitude to our HOD Ram Mohan Babu and to our guide Prof. Dr. Rajeshwari J, Associate Professor, Dayanand Sagar College of Engineering, for co-operating and guiding us through the process.

REFERENCES

[1] "Random Forests for Diabetes Diagnosis" by Sofia Benbelkacem and Baghdad Atmani, Laboratoire d'Informatique d'Oran (LIO) University of Oran 1 Ahmed Benbella Oran, Algeria. April 2019 International Conference on Computer and Information Sciences (ICCIS)

[2] "Classification of Diabetes Disease Using Support Vector Machine" by V. Anuja Kumari, R. Chitra Noorul Islam university. ISSN: 2248- 9622 www.ijera.com Vol. 3, Issue 2, March - April 2013, pp.1797-1801

[3] "Diabetes Prediction Based on XGBoost Algorithm" by Mingqi Li, Xiaoyang Fu and Dongdong Li. 2020 IOP Conf. Ser.: Mater. Sci. Eng. 768 072093

[4] "Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm" by Krati Saxena1, Dr. Zubair Khan, Shefali Singh. Volume 2 Issue 4, July-Aug 2014

[5] "Diabetics Prediction using Gradient Boosted Classifier" by J. Beschi Raja, R. Anitha, R.Sujatha, V. Roopa, S. Sam Peter. ISSN: 2249 – 8958, Volume-9 Issue-1, October 2019

[6] "Prediction of Diabetes by Using Artificial Neural Network, Logistic Regression Statistical Model and Combination of Them" by Paratoo RAHIMLOO, Ahmad JAFARIAN. Vol. 85, 2016, p. 1148 – 1164

[7] "Analysis and Detection of Diabetes using Data Mining Techniques- A Big Data application in health care" by B. G. Mamatha Bai, B. M. Nalini and Jharna Majumdar. Nitte Meenakshi Institute of Technology, January 2019