iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.62637

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

DiaVision - Diabetes Forecast using Machine Learning

Siddharth Kaushik[1], Aryan Gupta[2], Mohit Chauhan[3], Amit[4]
Computer Science Department, IMS Engineering College, Ghaziabad

Abstract: ***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 is a disease caused due to the increase level of blood glucose. This high blood glucose produces the symptoms of frequent urination, increased thirst, and increased hunger. Diabetes is a one of the leading causes of blindness, kidney failure, amputations, heart failure and stroke. When we eat, our body turns food into sugars, or glucose. At that point, our pancreas is supposed to release insulin. Insulin serves as a key to open our cells, to allow the glucose to enter and allow us to use the glucose for energy. But with diabetes, this system does not work. Type 1 and type 2 diabetes are the most common forms of the disease, but there are also other kinds, such as gestational diabetes, which occurs during pregnancy, as well as other forms. 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. The algorithms like K nearest neighbour, Logistic Regression, Random Forest, Support vector machine and Decision tree are used. The accuracy of the model using each of the algorithms is calculated. Then the one with a good accuracy is taken as the model for predicting the diabetes.***

I. INTRODUCTION

Diabetes is the fast-growing disease among the people even among the youngsters. In understanding diabetes and how it develops, we need to understand what happens in the body without diabetes. Sugar (glucose) comes from the foods that we eat, specifically carbohydrate foods. Carbohydrate foods provide our body with its main energy source everybody, even those people with diabetes, needs carbohydrate. Carbohydrate foods include bread, cereal, pasta, rice, fruit, dairy products and vegetables (especially starchy vegetables). When we eat these foods, the body breaks them down into glucose. The glucose moves around the body in the bloodstream. Some of the glucose is taken to our brain to help us think clearly and function. The remainder of the glucose is taken to the cells of our body for energy and also to our liver, where it is stored as energy that is used later by the body. In order for the body to use glucose for energy, insulin is required. Insulin is a hormone that is produced by the beta cells in the pancreas. Insulin works like a key to a door. Insulin attaches itself to doors on the cell, opening the door to allow glucose to move from the blood stream, through the door, and into the cell. If the pancreas is not able to produce enough insulin (insulin deficiency) or if the body cannot use the insulin it produces (insulin resistance), glucose builds up in the bloodstream (hyperglycaemia) and diabetes develops. Diabetes Mellitus means high levels of sugar (glucose) in the blood stream and in the urine.

Symptoms of Diabetes

1) Frequent Urination
2) Increased thirst
3) Tired/Sleepiness
4) Weight loss
5) Blurred vision
6) Mood swings
7) Confusion and difficulty concentrating
8) frequent infections

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, Coxsackievirus, mumps, hepatitis B virus, and cytomegalovirus increase the risk of developing diabetes.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

II. LITERATURE REVIEW

Recent advancements in technology, particularly in the realms of Internet of Things (IoT), Artificial Intelligence (AI), and machine learning (ML), have ushered in a transformative era in healthcare. AI, in particular, has revolutionized traditional healthcare management by enabling precise data-driven targeted care. This shift is particularly notable in the context of diabetes care, a field where early detection and effective management are paramount. The integration of ML techniques, has significantly enhanced the landscape of diabetes care. These methodologies play crucial roles in improving detection, prediction, and management resources for diabetes, offering the potential to identify patterns and trends that may not be apparent through traditional methods.

Furthermore, the availability of numerous tools and technologies within the ML and AI domains has facilitated automation in diabetes-related processes. For instance, predictive analytics models powered by ML algorithms can analyse vast datasets to identify risk factors and predict the likelihood of diabetes onset in individuals. This not only streamlines the diagnostic process but also enables healthcare providers to implement proactive interventions, such as lifestyle modifications or early treatment, to mitigate the progression of the disease.

The synthesis of these technologies underscores the importance of leveraging cutting-edge techniques to address the challenges posed by diabetes. By harnessing the power of AI and ML, healthcare professionals can gain deeper insights into patient data, personalize treatment plans, and optimize healthcare delivery. Moreover, the continuous evolution of these technologies offers promising avenues for further advancements in diabetes care, paving the way for more efficient, accurate, and patient-centred approaches to managing this complex metabolic disorder.

III. METHODOLOGY

In this section we shall learn about the various classifiers used in machine learning to predict diabetes. We shall also explain our proposed methodology to improve the accuracy. Five different methods were used in this paper. The different methods used are defined below. The output is the accuracy metrics of the machine learning models. Then, the model can be used in prediction.

A. Dataset Description

The diabetes data set was originated from https://www.kaggle.com/johndasilva/diabetes. Diabetes dataset containing 2000 cases. The objective is to predict based on the measures to predict if the patient is diabetic or not.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	2	138	62	35	0	33.6	0.127	47	1
1	0	84	82	31	125	38.2	0.233	23	0
2	0	145	0	0	0	44.2	0.630	31	1
3	0	135	68	42	250	42.3	0.365	24	1
4	1	139	62	41	480	40.7	0.536	21	0

Fig 1: Dataset Description

1) The diabetes data set consists of 2000 data points, with 9 features each.
2) "Outcome" is the feature we are going to predict, 0 means No diabetes, 1 means diabetes.
3) There are no null values in dataset.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

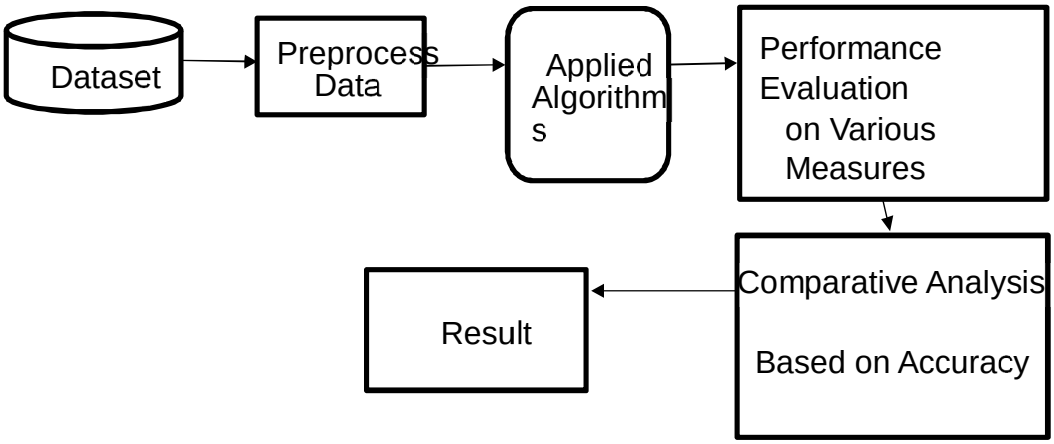


Fig 2: Proposed Model Diagram

IV. RESULT & DISCUSSION

A. *Correlation Matrix:*

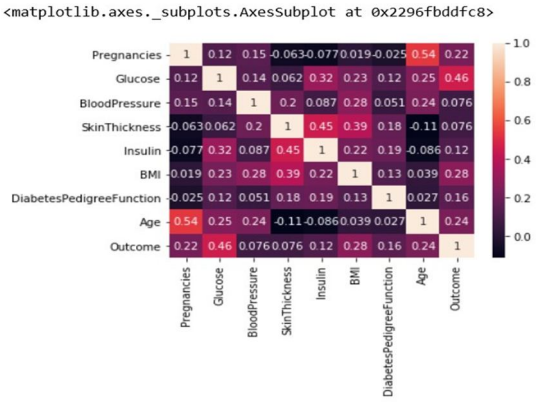


Fig 3: Correlation Matrix

It is easy to see that there is no single feature that has a very high correlation with our outcome value. Some of the features have a negative correlation with the outcome value and some have positive.

B. *Histogram:*

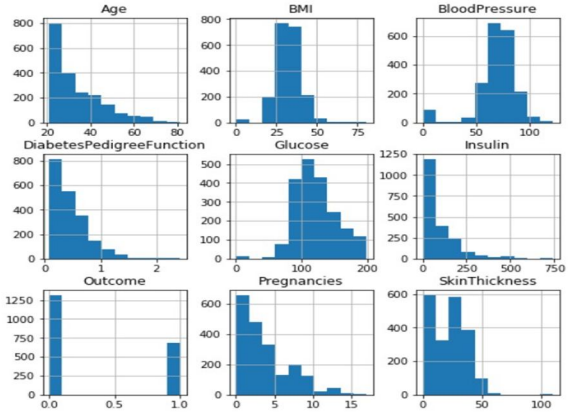
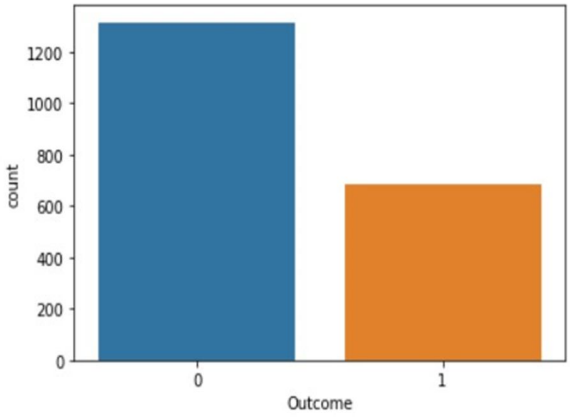


Fig 4: Histogram

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Let's take a look at the plots. It shows how each feature and label is distributed along different ranges, which further confirms the need for scaling. Next, wherever you see discrete bars, it basically means that each of these is actually a categorical variable. We will need to handle these categorical variables before applying Machine Learning. Our outcome labels have two classes, 0 for no disease and 1 for disease.

C. *Bar Plot for Outcome Class*

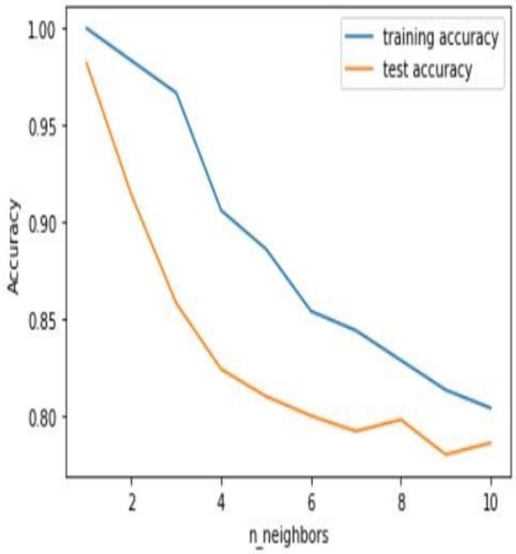


The above graph shows that the data is biased towards datapoints having outcome value as 0 where it means that diabetes was not present actually. The number of non-diabetics is almost twice the number of diabetic patients.

k-Nearest Neighbours:

The k-NN algorithm is arguably the simplest machine learning algorithm. Building the model consists only of storing the training data set. To make a prediction for a new data point, the algorithm finds the closest data points in the training data set, its "nearest neighbours."

First, let's investigate whether we can confirm the connection between model complexity and accuracy:



The above plot shows the training and test set accuracy on the y-axis against the setting of n-neighbours on the x-axis. Considering if we choose one single nearest neighbour, the prediction on the training set is perfect. But when more neighbours are considered, the training accuracy drops, indicating that using the single nearest neighbour leads to a model that is too complex. The best performance is somewhere around 9 neighbours.

Training Accuracy	0.78
Testing Accuracy	0.74

Table-1

D. Logistic regression:

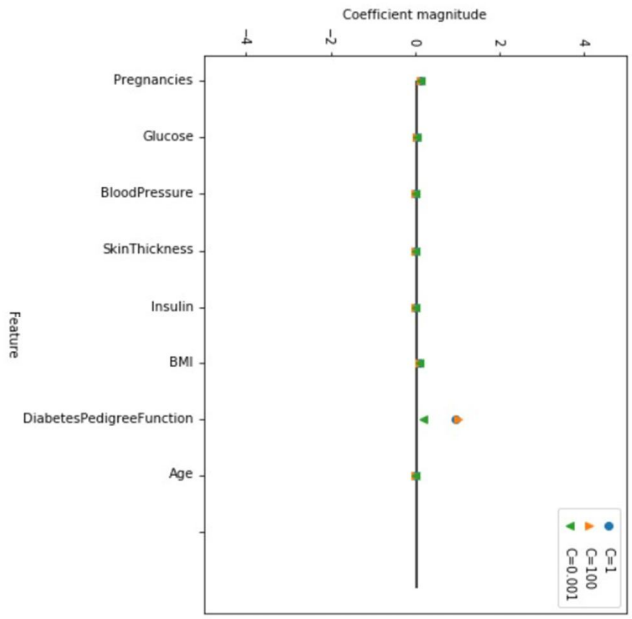
Logistic Regression is one of the most common classification algorithms.

	Training Accuracy	Testing Accuracy
C=1	0.779	0.788
C=0.01	0.784	0.780
C=100	0.778	0.792

Table-2

1) In first row, the default value of C=1 provides with 77% accuracy on the training and 78% accuracy on the test set.
2) In second row, using C=0.01 results are 78% accuracy on both the training and the test sets.
3) Using C=100 results in a little bit lower accuracy on the training set and little bit highest accuracy on the test set, confirming that less regularization and a more complex model may not generalize better than default setting.

Therefore, we should choose default value C=1.



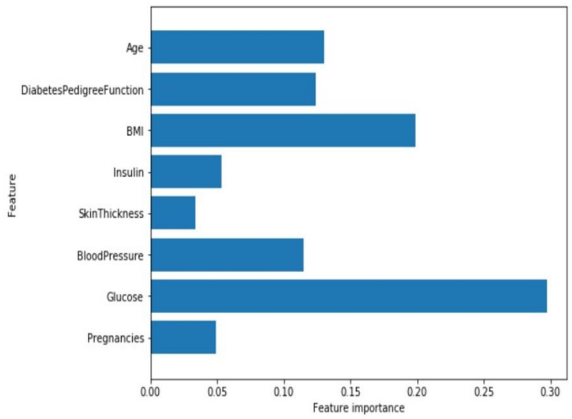
E. Decision Tree:

This classifier creates a decision tree based on which, it assigns the class values to each data point. Here, we can vary the maximum number of features to be considered while creating the model.

Training Accuracy	1.0
Testing Accuracy	0.74

Table-3

The accuracy on the training set is 100% and the test set accuracy is also good.
Feature Importance in Decision Trees
Feature importance rates how important each feature is for the decision a tree makes. It is a number between 0 and 1 for each feature, where 0 means “not used at all” and 1 means “perfectly predicts the target”.



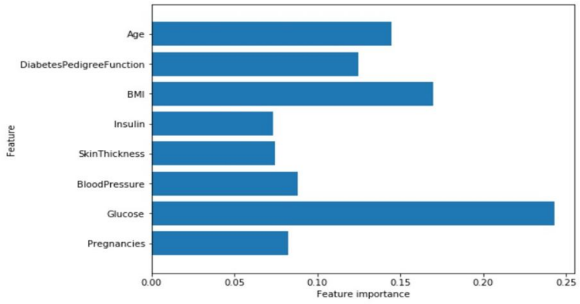
Feature “Glucose” is by far the most important feature.

F. Random Forest:
This classifier takes the concept of decision trees to the next level. It creates a forest of trees where each tree is formed by a random selection of features from the total features.

Training Accuracy	1.0
Testing Accuracy	0.74

Table-4

G. Feature importance in Random Forest:



Similarly to the single decision tree, the random forest also gives a lot of importance to the “Glucose” feature, but it also chooses “BMI” to be the 2nd most informative feature overall.

H. Accuracy Comparison:

Algorithms	Training Accuracy	Testing Accuracy
k-Nearest Neighbours	78.9%	74.67%
Logistic Regression	78%	74.52%
Decision Tree	98%	74.02%
Random Forest	98%	74.02%

Table-5

Table-5 shows the accuracy values for all machine learning algorithms.
Table-5 shows that k-Nearest Neighbours algorithm gives the overall best accuracy with 78.9% training accuracy and 74.67% Testing accuracy

V. CONCLUSION & FUTURE WORK

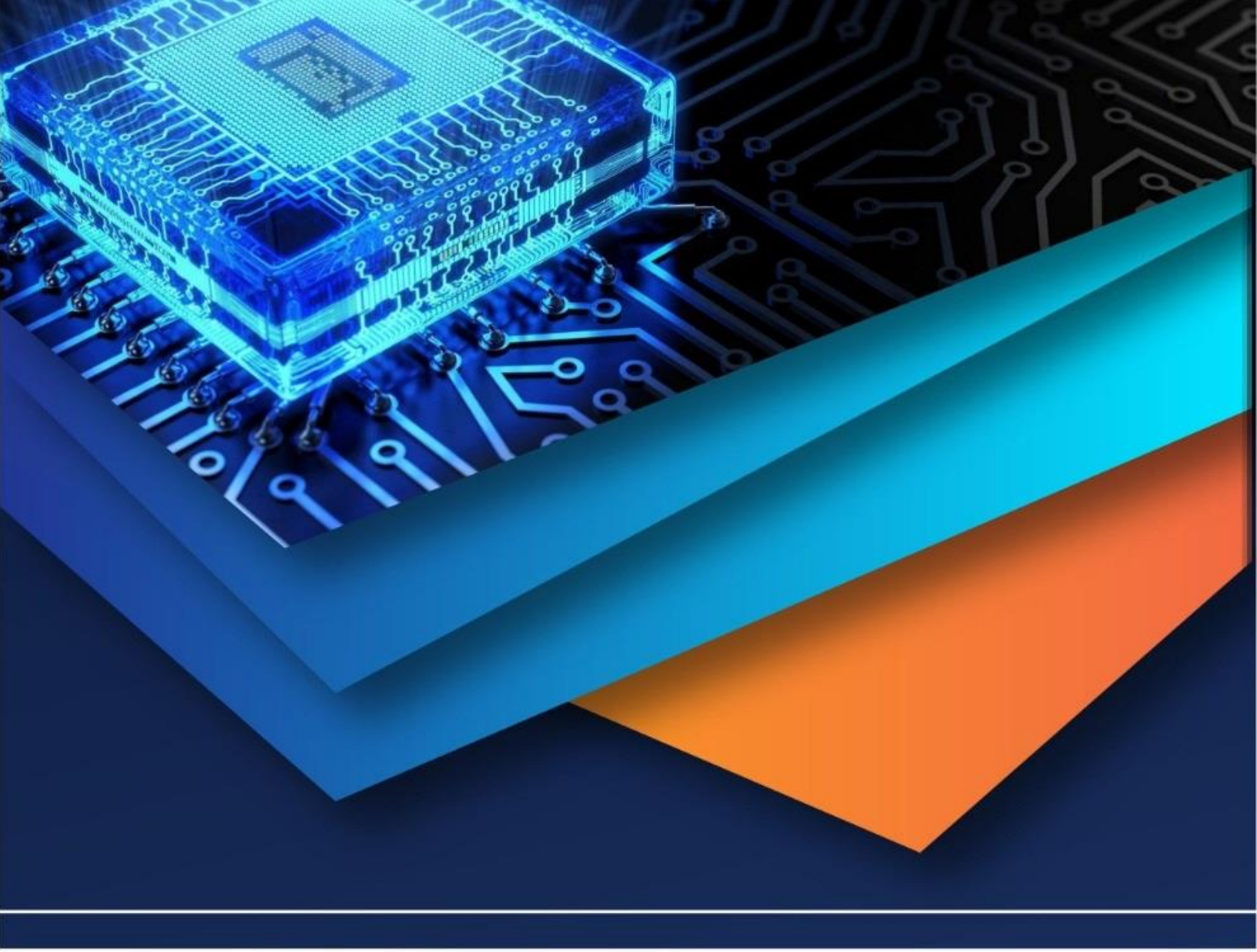
One of the important real-world medical problems is the detection of diabetes at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of diabetes. During this work, five machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on john Diabetes Database. Experimental results determine the adequacy of the designed system with an achieved accuracy of 78.9% using k-Nearest Neighbours algorithm.
In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.

VI. ACKNOWLEDGEMENT

We have completed this work under the mentorship of Dr. Pankaj Agarwal (Professor & Head) & Mr. Amit (Assistant Professor), Department of Computer Science & Engineering at IMS Engineering College, Ghaziabad. We are doing an online summer internship on Machine Learning where I have learnt the various Machine Learning Algorithms from both of my mentors as Course Instructors. This work is been assigned as project assignments to us.
We would like to express my special thanks to both of my mentors for inspiring us to complete the work & write this paper. Without their active guidance, help, cooperation & encouragement, we would not have our headway in writing this paper. we are extremely thankful for their valuable guidance and support on completion of this paper.
We extend our gratitude to “IMS Engineering College” for giving us this opportunity. We also acknowledge with a deep sense of reverence, our gratitude towards our parents and member of my family, who has always supported us morally as well as economically. Any omission in this brief acknowledgement does not mean lack of gratitude.

REFERENCES

[1] Isfafuzzaman Tasin, Tansin Ullah Nabil, Sanjida Islam and Riasat Khan. Diabetes prediction using machine learning and explainable AI techniques doi: 10.1049/ht12.12039
[2] Arwatki Chen Lyngdoh, Nurul Amin Choudhury and Soumen Moulik. Diabetes Disease Prediction Using Machine Learning Algorithms, 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES) - IECBES 2020 DOI: IO. 1109/lECBES48179.2021.9398759
[3] S. Rakesh Kumar, Kruthi. G and V. Supraja. Diabetes Prediction with Machine Learning with Python, March 2024 International Journal of Scientific Research in Computer Science Engineering and Information Technology 10(2):100-106 DOI: 10.32628/CSElT2390651
[4] Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., 2013. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University - Computer and Information Sciences 25, 127–136. doi:10.1016/j.jksuci. 2012.10.003
[5] Deepti Sisodia and Dilip Singh Sisodia. Prediction of Diabetes using Classification Algorithms National Institute of Technology, G.E Road, Raipur and 492001, India
[6] B. Dhomse Kanchan and Kishor M. Mahale. Study of machine learning algorithms for special disease prediction using principal of component analysis Conference: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), DOI: 10.1109/lCGTSPlCC.2016.7955260



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY