IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 5, May 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005221|

Diabetes Prediction Using Machine Learning

Chimakurthy Kavya Sri[1], Chinnam Abhiram[2], B.Ganesh Varma[3], Akula Abhishek[4], D. Deepthi[5]
U.G.Student, Department of Computer Science and Engineering, Vasireddy Venkatadri Institute of Technology, Nambur, Andhra Pradesh, India[1,2,3,4]
Associate professor, Department of Computer Science and Engineering, Vasireddy Venkatadri Institute of Technology, Nambur, Andhra Pradesh, India[5]

ABSTRACT: One of the most deadly and debilitating disorders associated with elevated blood sugar levels is diabetes. If diabetes isn't treated and diagnosed, it can lead to a slew of problems. As a result of the lengthy identification process, patients end up visiting diagnostic centres and seeing doctors. Machine learning, on the other hand, provides a solution to this essential issue. With this research, the researchers are hoping to create a model that can accurately predict whether or not a patient will develop diabetes. A person develops diabetes when their blood glucose levels rise to dangerously high levels. If left untreated, diabetes can lead to serious health concerns, including heart disease, kidney disease, high blood pressure, eye damage, and other organ damage. If diabetes is detected early enough, it can be managed. In order to accomplish this goal, we will use several machine learning techniques to forecast diabetes in a human body or a patient at an early stage. Techniques of machine learning Improve prediction accuracy by building models from patient-generated data. Machine Learning Classification and ensemble methods will be used to predict diabetes in this study.. A decision tree, a support vector machine (SVM), a gradient booster (GB), and a random forest are all examples of K-Nearest Neighbor (RF). Every model has a varied level of accuracy when compared to others.

I.INTRODUCTION

Diabetes is one of the most dangerous diseases there is. Obesity, elevated blood glucose levels, and other factors can all contribute to diabetes. Diabetes-related abnormalities in crab metabolism are alleviated, and blood sugar levels are brought back to normal. When the body does not produce enough insulin, it results in diabetes. Diabetes affects 422 million people worldwide, with the majority living in countries with poor or no income, according to the World Health Organization (WHO). And by 2030, this might rise to 490 billion dollars. However, many countries, such as Canada, China, and India, have a high prevalence of diabetes. There are currently more over 100 million people in India, hence the true number of diabetics in the country is 40 million. Diabetic complications are a leading cause of death worldwide. Diabetes can be prevented and even cured if diagnosed early enough. For this, this study investigates the prediction of diabetes by taking into account numerous diabetes-related features. The Pima Indian Diabetes Dataset is used in conjunction with a variety of machine learning classification and ensemble techniques in order to make this prediction. Computers or machines can be trained using Machine Learning. Various machine learning techniques provide efficient results for collecting knowledge by creating various classification and ensemble models from gathered datasets. They are. Diabetes can be predicted using these information. Predictions can be made using a variety of Machine Learning approaches, but selecting the right one is difficult. Hence the common classification and ensemble approaches are used for this purpose.

II.LITERATURE SURVEY

2.1 Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance's for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018

Diabetes is an uprising illness, particularly because of the kind of nourishment we are having these days and the conflicting eating regimen and schedule that we take after. Diabetes are fundamentally caused because of obesity or high glucose level, and so forth. So in this paper we will discover what are the critical elements for the reason for diabetes. Variable and feature choice have turned into the focal point of much research in regions of utilization for which datasets with tens or a huge number of factors are accessible.. Likewise we will center around the most essential features to predict whether a person will have chances to develop diabetes in the future

Diabetes is an uprising illness, particularly because of the kind of nourishment we are having these days and the conflicting eating regimen and schedule that we take after. Diabetes are fundamentally caused because of obesity or high glucose level, and so forth. So in this paper we will discover what are the critical elements for the reason for diabetes. Variable and feature choice have turned into the focal point of much research in regions of utilization for which datasets with tens or a huge number of factors are accessible.. Likewise we will center around the most essential features to predict whether a person will have chances to develop diabetes in the future.

Diabetes mellitus has an immediate flag of high glucose, together with a few side effects including continuous pee, expanded thirst, expanded yearning and weight reduction. Patient of diabetes for the most part require consistent treatment, else, it will potentially prompt numerous perilous hazardous complications. The diabetes is determined to have the 2-hour post-stack plasma glucose being no less than 200mg/dL [1], and the need of recognizing diabetes convenient brings in different examinations about diabetes acknowledgment. Numerous past research thinks about have been done about machine learning in diabetes recognizable proof.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005221|

Research has been done centered around diabetes recognizable proof through SVM (Support Vector Machine) [2] and they acquired some rousing outcomes. Contrasting with the past work, we make a more comprehensive examination containing various regular systems used to diabetes ID, proposing to think about their execution and locate the best one among them. Through this investigation, we look at a few normal and information preprocessors for every one of the classifiers we utilize, and locate the best preprocessor separately. At that point we think about using different classifiers after we adjust the parameters and different kernels of them to achieve their surmised most extreme precision.. Finally, we additionally investigate the pertinence of each element with the arrangement result, and this will change the informational index in future examinations.

2.2 Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019

In people with diabetes mellitus, elevated blood sugar levels persist for an extended length of time as a result of a variety of metabolic abnormalities. It affects a wide range of bodily systems, including the blood vessels and nerves, and hence has a devastating effect on the human body as a whole. It is possible to regulate and save human lives by doing early disease prediction. Using machine learning techniques, this study aims to identify distinct risk factors for this condition. Using machine learning approaches, diabetic patients' diagnostic medical records can be used to build predictive models that can be used to extract knowledge. Predicting diabetic patients can be aided by the extraction of knowledge from such data. On the basis of data from the adult population and four common machine learning methods (SVM, Naive Bayes, K-Nearest Neighbor, and C4.5 Decision Tree) we can predict diabetes. The C4.5 decision tree performed better than other machine learning algorithms in our experiments.

III.PROPOSED SYSTEM

A person develops diabetes when their blood glucose levels rise to dangerously high levels. If left untreated, diabetes can lead to serious health concerns, including heart disease, kidney disease, high blood pressure, eye damage, and other organ damage. If diabetes is detected early enough, it can be managed. In order to accomplish this goal, we will use several machine learning techniques to forecast diabetes in a human body or a patient at an early stage. Techniques of machine learning Improve prediction accuracy by building models from patient-generated data. Machine Learning Classification and ensemble methods will be used to predict diabetes in this study.. A decision tree, a support vector machine (SVM), a gradient booster (GB), and a random forest are all examples of K-Nearest Neighbor (RF). Every model has a varied level of accuracy when compared to others.

3.1 Dataset Description

This study made use of the widely-used Pima Indian diabetes dataset . The Pima Indian, a native American population in Phoenix Arizona, USA, has been continually studied and examined since 1965 due to the high incidence of diabetes . Pima Indian females participated in standardized diabetes examinations. Three additional datasets were derived from the Pima Indian dataset and used in this study.

In this dataset, diabetes was diagnosed according to the World Health Organization criteria
. A patient was considered diabetic if the 2 hour post-load plasma glucose was at least 200
mg/d1(11.1mmol/l) at any survey examination, or if the Pima Indian Health Service Hospital serving the community found a glucose concentration of at least 200 mg/d1 during the course of routine medical care .

The Pima Indian dataset has 768 observations, each with nine attributes.Five hundred of these patients were non diabetic, and the rest were diabetic. Each patient has only one record in the dataset. The attributes in this dataset are given below. Basic statistical properties of the dataset are

1. Number of times pregnant (PREG)
2. Plasma Glucose Concentration at 2 Hours in an Oral Glucose Tolerance Test (GLUC)
3. Diastolic Blood Pressure (mmHg) (PRESS)
4. Triceps Skin Fold Thickness (mm) (SKIN)
5. 2-Hour Serum Insulin (Uh/ml) (INSU)
6. Body Mass Index (Weight in kg / (Height in cm)) (BMI)
7. Diabetes Pedigree Function (PDF)
8. Age in years (AGE)
9. Diabetes Class Variable (0 or 1)

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

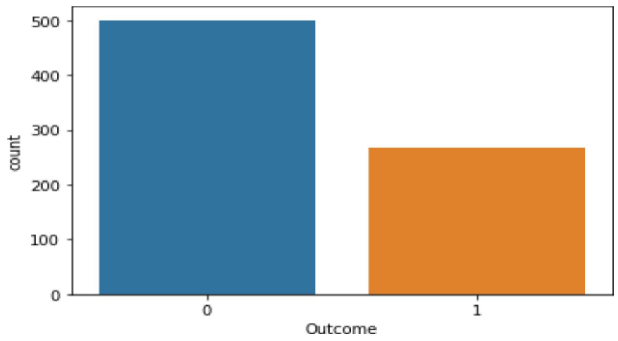
|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005221|

Feature	Minimum	Maximum	Mean	Variance	Standard Deviation
Number of times pregnant	0.00	17.00	3.85	11.34	3.37
Plasma Glucose Concentration at 2 Hours in an Oral Glucose Tolerance Test	0.00	199.00	120.89	1020.92	31.95
Diastolic Blood Pressure (*mmHg*)	0.00	122	69.11	374.16	19.34
Triceps Skin Fold Thickness (*mm*)	0.00	99.00	20.54	254.14	15.94
2-Hour Serum Insulin (*Uh/ml*)	0.00	846	79.8	13263.89	115.17
Body Mass Index (Weight in *kg* / (Height in *cm*))	0.00	67.10	31.99	62.08	7.88
Diabetes Pedigree Function	0.08	2.42	0.47	0.11	0.33
Age (years)	21.00	81.00	33.22	138.12	11.75

Table-Statistical properties of the dataset

Distribution of Diabetic patient- We made a model to predict diabetes however the dataset was slightly imbal- anced having around 500 classes labeled as 0 means negative means no diabetes and 268 labeled as 1 means positive means diabetic.



3.2 DATA PREPROCESSING

All attribute values in the Pima Indian dataset were re-scaled (normalized) to avoid the disadvantages with using data that has varying value scales for classification. For example,
age and blood pressure attributes are on different numerical scales. Min-Max, Z-score, and Decimal Scaling are popular normalization techniques.

In this study, the Min-Max Normalization technique was used to re-scale all attribute values to be in the range [0,1]. This technique was selected because of its simplicity and ability to preserve the relationships among the original data values. Min-Max Normalization provides a linear transformation on the original range of data by computing a new value for each attribute of an instance.

1. Remove all instances of zero (0) as a value and discard them. It's impossible to be worth nothing. As a result, we can disregard this instance. Feature subset selection, which lowers the diamentonality of data and speeds up labour, is accomplished by removing irrelevant characteristics and instances.
2. The data is split into training and testing sets after it has been cleaned. When data is spit out, we use the training data set to train the algorithm and the test data set for testing purposes only. The training model will be built using logic and algorithms, as well as the values of features in the training data, in this process. To put it simply, the goal of normalisation is to bring all of the qualities to the same level.

3.3 Apply Machine Learning

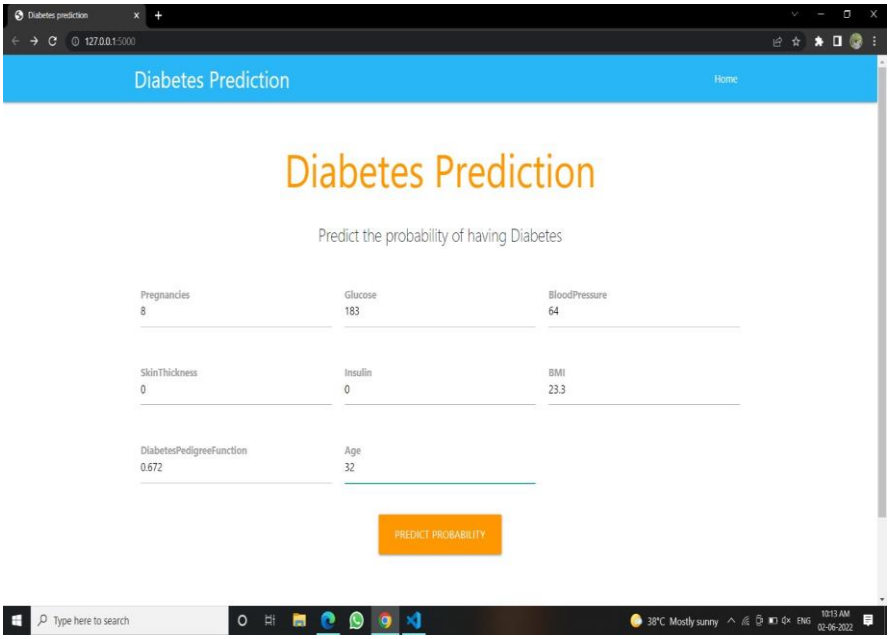
When data has been ready we apply Machine Learning Technique. We use different classification and ensemble techniques, to predict diabetes. The methods applied on Pima Indians diabetes dataset. Main objective to apply Machine Learning Techniques to analyze the performance of these methods and find accuracy of them, and also been able to figure out the responsible/important feature which play a major role in prediction.

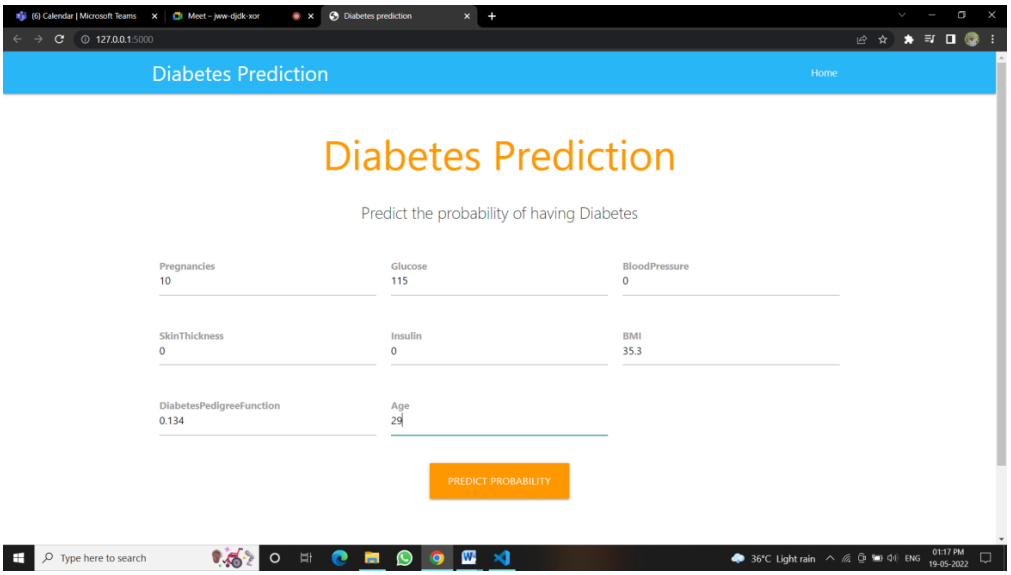
| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

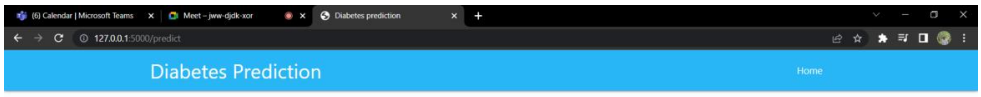
|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005221|

IV.RESULTS AND DISCUSSION





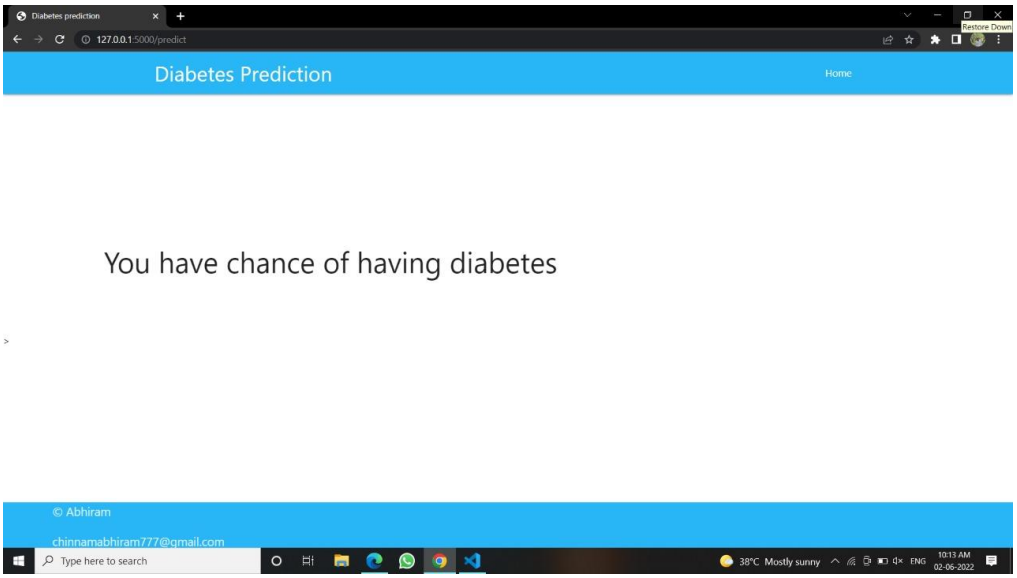


You are safe

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005221|



V.CONCLUSION

In this work, six upgraded Machine Learning approaches were used to address the Type II diabetes diagnosis problem. A Statistical Regression model and instance-based techniques were used to tackle the type II diabetes diagnosis problem using three different forms of upgraded Artificial Neural Nets. Type II diabetes was diagnosed using these methods on a variety of datasets, and the results were compared to those of other studies in the literature. Many of the machine learning strategies proposed in this study performed as well as or better than those found elsewhere in the literature.

To train and assess machine learning algorithms described in this paper, the Pima Indian diabetes dataset was preprocessed using several data preprocessing and feature extraction methods. Techniques like Min-Max normalisation and AHP were used in the data preprocessing. Preprocessed datasets demonstrated a considerable improvement in the ML approaches in this study. On the original dataset with missing values and the dataset with imputed data, most of these approaches failed to reach their optimal performance. Compared to datasets where instances with missing values were eliminated or important characteristics were identified, the proposed strategies yielded worse results on these two datasets. This means that data imputation is not as effective as feature extraction and the removal of instances with missing values or a combination of both.

According to a literature study, most research solely employed classification accuracy to demonstrate the usefulness of their Machine Learning procedures, rather than adding additional performance measures to clearly indicate the efficacy of their Machine Learning approaches. When it comes to diagnosing type II diabetes, ML approaches have been proven to be effective by this study's use of several performance measures. Comparing ANNs against Instance-based and Statistical approaches, this study found that ANNs performed the best. With 81 percent classification accuracy, 89 percent sensitivity, and 88 percent specificity, the Instance-based approaches had the greatest results. In terms of classification accuracy, the statistical method achieved a figure of 84% with a sensitivity of 72% and a specificity of 100%. In the absence of preprocessing, the ANNs performed admirably, as they did when combined with feature extraction was performed to fill up the blanks. Instance-based approaches performed best when the dataset had no missing values. When missing values were eliminated and attribute means were imputed into the dataset, the statistical approach performed at its best.

In this study and prior investigations, it can be stated that Artificial Neural Networks are more effective in diagnosing type II diabetes than other Machine Learning methods. Data pretreatment approaches such as feature extraction can also be used to improve the performance of ML algorithms by eliminating instances with missing values. For the most part, the results of this study reveal that novel approaches, hybridization, and efficient preparation of the dataset improved the performance of ML techniques in diagnosing type II diabetes, with one of the best results being a classification accuracy of 100 percent.

Using the proposed data preparation techniques on datasets from diverse domains could be an avenue for future research in this area. Performing a temporal complexity analysis on the proposed methodologies in this paper would be valuable in the future. The proposed machine learning techniques could be used in other domains to see whether better results can be reached in the future.

REFERENCES

[1] World Diabetes Foundation, International Diabetes Federation, Novo Nordisk, Diabetes South Africa, Diabetes: the hidden pandemic and its impact on Sub-Saharan Africa

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005221|

- document prepared for the Diabetes Leadership Forum Africa 2010 - (Edited by Prof Ayesha Motala and Dr Kaushik Ramaiya, 2010).

[2] The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), Causes of diabetes, Website: http://diabetes.niddk.nih.gov/dm/pubs/causes/, Accessed: 2015/02/08.

[3] American Diabetes Association, Diagnosis and Classification of Diabetes Mellitus, Website: http://www.ncbi.mm.nih.gov/pmc/articles/PMC2613584/, Accessed: 2015/02/08.

[4] International Diabetes Federation, Follow-up to the political declaration of the high- level meeting of the General Assembly on the prevention and control of non-communicable diseases, IDF Diabetes Atlas 6th Edition, 2013.

[5] Diabetes Atlas 6th Edition Committee, IDF Diabetes Atlas 6th Edition Poster Update, 2014.

[6] Shankaracharya, Devang Odedra, Medhavi Mallick, Prateek Shukla, Subir Samanta, and Ambarish Vidyarthi, Java-Based Diabetes Type 2 Prediction Tool for Better Diagnosis, Diabetes technology and therapeutics, Vol- ume 14, Number 3, 2012.

[7] Nahla H. Barakat, Andrew P. Bradley, and Mohamed Nabil H. Barakat, Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus, IEEE transactions on information technology in biomedicine, Vol. 14, Number 4, July 2010.

[8] Krati Saxena, Zubair Khan, and Shefali Singh, Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algo- rithm, International Journal of Computer Science Trends and Technology (IJCST) - Volume 2 Issue 4, July-Aug 2014.

[9] N. Lavrac, E. Keravnou, and B. Zupan, Intelligent data analysis in medicine, Encyclopedia of Computer Science and Technology, Vol. 42, 2000, pp. 113-157.

Najmeh Hosseinpour, Saeed Setayeshi, Karim Ansari-asl and Mohammad Mosleh, Diabetes Diagnosis by Using Computational Intelligence Algorithms, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 12, December 2012

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details