International
Journal
Of Advanced Research in Engineering & Management (IJAREM)
ISSN: 2456-2033 || PP. 01-06

Diabetes Prediction using Machine Learning

Shaurya Goel
Computer Engineering Department, Thapar University

Dr. Saurabh Goel (Professor), Abhinav Anand, Dharmbir, Bhavika[#]
CSE Department, Panipat Institute of Engineering and Technology

Abstract: Diabetes is a chronic illness that could lead to a global health crisis. The International Diabetes Federation estimates that 382 million people worldwide have diabetes. This will double to 592 million by 2035. Diabetes is a condition brought on by elevated blood glucose levels. The symptoms of increased thirst, increased appetite, and frequent urination are brought on by this elevated blood glucose. Diabetes is one of the main causes of heart failure, stroke, amputations, kidney failure, and blindness. The two most prevalent types of diabetes are type 1 and type 2, but there are other varieties as well, including gestational diabetes, which develops during pregnancy. In data science, machine learning is a young scientific discipline that studies how machines pick up knowledge via experience. The goal of this project is to combine the findings of many machine learning approaches to create a system that can more accurately forecast a patient's risk of developing diabetes at an early age.

Keywords: Random Forest, machine learning, diabetes, and accuracy, corelation matrix, confusion matrix.

I. Introduction

Even in young people, diabetes is a disease that is rapidly spreading throughout society.

We must comprehend what occurs in the body without diabetes if we are to comprehend diabetes and how it arises. We get sugar, or glucose, from the meals we eat—more especially, from diets high in carbohydrates.

Our bodies need carbohydrates as their primary energy source, so everyone—even those who have diabetes—needs them. Bread, cereal, pasta, rice, fruit, dairy products, and vegetables—especially starchy vegetables—are examples of foods high in carbohydrates. The body converts these meals into glucose when we eat them. The bloodstream carries the glucose throughout the body. A portion of the glucose is transported to the brain to support proper brain function. The remaining glucose is absorbed by our body and used as fuel by our cells. It is also stored in our liver for use by the body at a later time. Insulin is needed so that the body can use glucose as fuel. The beta cells in the pancreas create the hormone insulin. Insulin functions similarly to a door key. Insulin binds itself to cell doors, allowing them to let bloodstream glucose enter the cell through the door.

Types of Diabetes:

Type 1- A weakened immune system and insufficient insulin production by the cells are the hallmarks of type 1 diabetes. There are currently no effective preventative measures for type 1 diabetes, nor are there compelling research to support its aetiology.

Type 2- This type of diabetes is characterised by either insufficient insulin production by the cells or improper insulin utilisation by the body. Ninety percent of people with diabetes are diagnosed with this kind of diabetes, making it the most prevalent type. It is brought on by both inherited traits and lifestyle choices.

When pregnant women have an abrupt rise in blood sugar, they may develop gestational diabetes. It will return in two thirds of the instances in subsequent pregnancies. Type 1 or type 2 diabetes is very likely to develop during a gestational diabetes-affected pregnancy.

Symptoms of Diabetes:

Frequent urge of Urination, Enhanced thirst, Tiredness or drowsiness, Loss of weight, Hazy vision, Changes in mood and Frequent infections causing confusion and difficulties concentrating

Causes of Diabetes:

The primary cause of diabetes is genetic. It is brought on by at least two mutated genes in chromosome 6, which controls how the body reacts to different antigens.

Infection with viruses can potentially affect the development of type 1 and type 2 diabetes. Research has indicated that viral infections, including cytomegalovirus, hepatitis B virus, mumps, rubella, and Coxsackievirus, elevate the likelihood of getting diabetes.

II. Literature Review

In order to determine if a person is diabetic or not, Yasodha et al. employ classification on a variety of datasets. The data set for the diabetic patient is created by compiling information from the hospital warehouse, which has 200 instances with 9 different attributes. The two groups that are being discussed here in this dataset are the blood tests and the urine tests. The implementation in this study can be carried out by classifying the data using WEKA, and the data is evaluated using the 10-fold cross-validation approach, which works incredibly well on tiny datasets, and comparing the results. The J48, Random, REP, and naïve Bayes models are employed. The results showed that, among the others, J48 performs the best, with an accuracy of 60.2%.

Aiswarya et al.'s goal is to find ways to identify diabetes by looking into and analysing the patterns that emerge in the data through classification analysis with the use of Decision Tree and Naïve Bayes algorithms. The goal of the research is to provide a quicker and more effective way to diagnose the condition, which would aid in the timely treatment of the patients. The study found that the J48 method yields an accuracy rate of 74.8% whereas the naïve Bayes algorithm yields an accuracy of 79.5% by adopting a 70:30 split, using the PIMA dataset and a cross-validation approach.

The study compares the performance of the same classifiers when implemented on some other tools, including Rapid-miner and Matlab using the same parameters (i.e. accuracy, sensitivity, and specificity). Gupta et al. aims to find and calculate the accuracy, sensitivity, and specificity percentage of numerous classification methods and also tried to compare and analyse the results of several classification methods in WEKA. They used the Bayes Net, Jgraft, and JRIP algorithms. According to the results, Jgraft has the highest accuracy (81.3%), sensitivity (59.7%), and specificity (81.4%). Additionally, it was determined that WEKA performs better than Matlab and Rapid Mine.

After running the resample filter over the data, Lee et al. concentrate on using the CART decision tree method on the diabetes dataset. In order to obtain higher accuracy rates, the author places attention on the class imbalance issue and the necessity of addressing it before implementing any algorithms. Class imbalances are typically found in datasets with dichotomous values, meaning that each class variable has two possible outcomes. If this is noticed early on in the data preprocessing stage, it can be easily handled and will improve the predictive model's accuracy.

III. Methodology

This part will teach us about the several classifiers that are used in machine learning to forecast diabetes. We will also go over our suggested methods for increasing accuracy. This paper employed five distinct approaches. The various techniques are described below. The machine learning models' accuracy measurements are the output. After then, predictions can be made using the model.

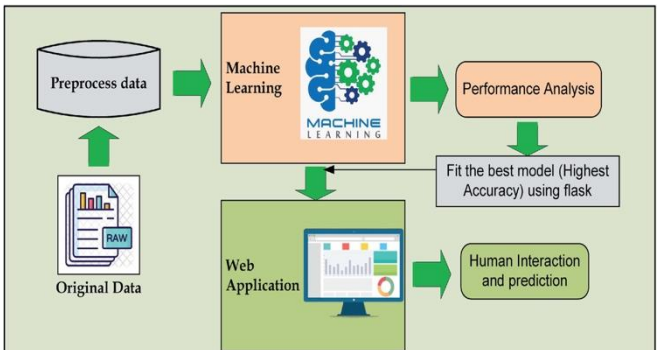
Description of the Dataset

The source of the diabetes data set was *https://www.kaggle.com/code/ahmetcankaraolan/diabetes-prediction-using-machine-learning/input*. Diabetes dataset of instances. The goal is to determine whether or not the patient has diabetes by making predictions based on the measurements.

	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

➔ "Outcome" is the feature we are going to predict. 0 means No diabetes, 1 means diabe

Table- This table details regarding our dataset of 768 entries.



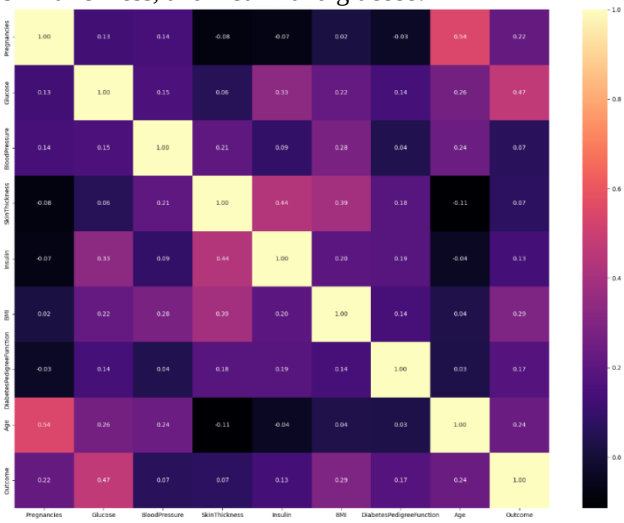
IV. Result and Discussions

One of the best datasets for testing machine learning algorithms for diabetes prediction is the Pima Indian Diabetes Dataset (UCI Machine Learning Repository, 1998). The Pima Indian dataset was made available by the National Institute of Diabetes and Digestive and Kidney Diseases. It uses diagnostic metrics such as age, BMI, skin thickness, blood pressure, insulin, glucose level, and diabetes pedigree function to assess if a patient has the disease.

Corelation Matrix:

The corelation matrix indicates the relationship between the features and the target variable. The correlation diagram shows how the parameters relate to one another. Age, BMI, glucose, and pregnancy status are the characteristics most closely linked to the outcome.

The results are unaffected by insulin and diabetes pedigree function; blood pressure and skin thickness have a weak correlation with the outcome; and there is a modest correlation between age and pregnancy, insulin and skin thickness, BMI and skin thickness, and insulin and glucose.



It is evident that no single attribute significantly correlates with our result value. Certain features exhibit a negative association with the outcome value, whereas others show a positive correlation.

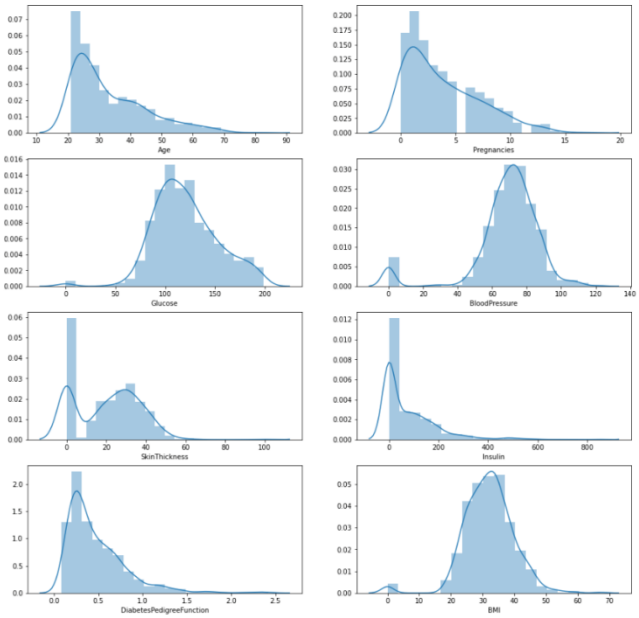
Confusion Matrix:

A table used in statistics and machine learning to evaluate a classification model's performance is called a confusion matrix. By displaying the counts of true positive, true negative, false positive, and false negative predictions, it provides an overview of the categorization findings.

```
array([[92,  9],
       [10, 41]], dtype=int64)
```

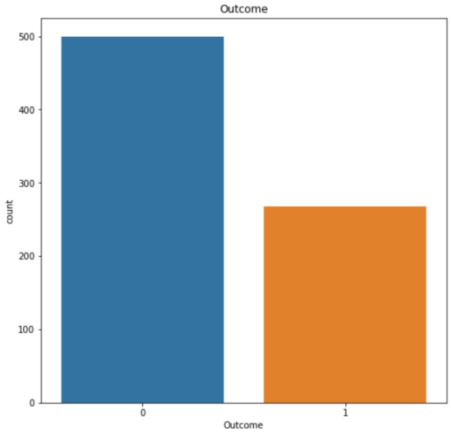
Fig- Confusion matrix of our dataseset

Histogram:



Let's examine the storylines. It provides additional evidence for the necessity of scaling by displaying the distribution of each feature and label over several ranges. Next, each discrete bar you see indicates that this is a categorical variable in and of itself. Prior to using machine learning, these categorical factors must be addressed. We have two classifications for our outcome labels: 0 for no disease and 1 for disease.

Bar Plot for Outcome Class:

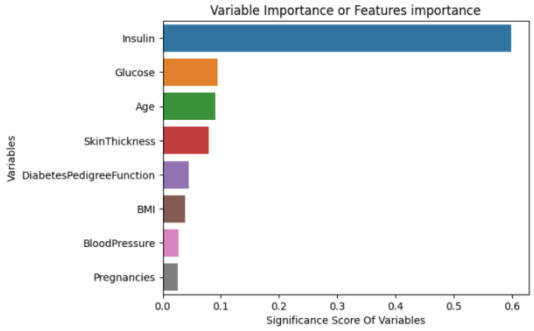


The graph above demonstrates how the data is skewed in favour of data points with an outcome value of 0, which indicates that diabetes was not genuinely present. The proportion of those without diabetes is about double that of those with the disease.

Random Forest:

The idea of decision trees is advanced by this classifier. It produces a forest of trees, using a random selection of features from the total features forming each tree.

Feature Importance in Random Forest:



V. Conclusion and Future Work

Early identification of diabetes is one of the major medical issues that arise in real life. The present work aims to develop a system that can anticipate the presence of diabetes through methodical efforts. In the course of this paper, five machine learning classification algorithms are examined and assessed using different metrics. Research is conducted using the John Diabetes Database. Based on experimental data, the Decision Tree algorithm is used to determine the suitability of the planned system with an accuracy of 99%.

VI. Acknowledgement

I have completed this work under the mentorship of Dr. Saurabh Goel , Department of Computer Science & Engineering at Panipat Institute of Engineering and Technology College, Samalkha, Haryana. My mentor, who is also my course instructor, is teaching me the different machine learning algorithms this work is been assigned as project assignments to us.

I want to thank my mentor in particular for motivating us to finish the assignment and produce this paper. We would not have made progress in drafting this work if it weren't for their enthusiastic direction, assistance, collaboration, and encouragement. I sincerely appreciate all of their helpful advice and assistance in finishing this assignment.

I would want to express my appreciation to Panipat Institute of Engineering and Technology College for providing me with this chance. I also express my sincere gratitude, filled with reverence, to my parents and other family members who have always provided me with both financial and moral assistance. Any information left out of this succinct acknowledgement does not imply a lack of appreciation.

VII. References

[1]. G. Parimala, R. Kayalvizhi, S. Nithiya, 2023 International Conference on Computer Communication and Informatics (ICCCI), doi:10.1109/ICCCI56745.2023.10128216

[2]. Nazin Ahmed, Rayhan Ahammed, Md. Manowarul Islam, Md. Ashraf Uddin, Arnisha Akhter, Md. Alamin Talukder, Bikash Kumar Paul, 2021, Machine learning based diabetes prediction and development of smart web application, doi: doi.org/10.1016/j.ijcce.2021.12.001

[3]. Choubey, D. K., Paul, S., Kumar, S., Kumar, S., 2017. Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection, in: Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016), pp. 451–455.

[4]. Rani, A. S., & Jyothi, S. (2016, March). Performance analysis of classification algorithms under different datasets. In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on (pp. 1584- 1589). IEEE

[5]. Dhomse Kanchan B., M.K.M., 2016. Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis, in: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE. pp. 5–10.

[6]. Sisodia, D., Singh, L., Sisodia, S., 2014. Fast and Accurate Face Recognition Using SVM and DCT, in: Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012, Springer. pp. 1027–1038.

[7]. Bamnote, M.P., G.R., 2014. Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Systems and Computing 1, 763–770. doi: 10.1007/978-3-319-11933-5.
[8]. Aljumah, A.A., Siddiqui, M.K., and Ahamad, M.G. (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.
[9]. Arora, R., Suman, 2012. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. International Journal of Computer Applications 54, 21–25. doi:10.5120/8626-2492.
[10]. https://www.kaggle.com/code/ahmetcankaraolan/diabetes-prediction-using-machine-learning/input.