IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Review on Diabetes Detection using machine learning techniques

Tarun Viswakarma, Yavan Mahilang, Yashwit Soni

Master of Computer Application

Medi-Caps University, Indore, India

Abstract: Diabetes mellitus is a chronic disease that occurs as a result of the pancreas not producing enough insulin or the body's inability to use the insulin it produces effectively. Insulin is a drug that controls blood sugar. A fasting blood sugar level of 70-110 mg/dL is considered normal, 100-125 mg/dL is considered diabetes, and 126 mg/dL and above is considered diabetes. The number of people with diabetes increased from 108 million in 1980 to 422 million in 2014. The rate of increase was faster in low- and middle-income countries than in high-income countries. There are two types of diabetes - type1 diabetes and type2 diabetes.

This section explains the causes of both types of diabetes. Diagnosis of diabetes: Many scientists and doctors are now developing artificial intelligence-based diagnostic methods to better solve problems caused by human error. KNN, Support Vector Machine (SVM), decision trees, Random Forest etc. Various types of machine learning are discussed and compared.

Keywords: **Diabetes mellitus, Pancreas, Insulin, Blood sugar, Fasting blood sugar level, Type1 diabetes, Type2 diabetes, Causes of diabetes, Diagnosis of diabetes, Artificial intelligence, KNN, Support Vector Machine (SVM), Decision trees, Random Forest, Machine learning.**

1. INTRODUCTION

Diabetes is a medical condition that arises when the level of glucose, commonly known as blood sugar, in the blood becomes too high. Glucose is the primary source of energy for the body and is obtained from the food we consume. The pancreas produces a hormone called insulin, which facilitates the absorption of glucose from the food into the cells for energy production. However, in some cases, the body does not produce enough insulin or cannot effectively use the insulin produced, leading to glucose accumulation in the blood and lack of energy production in the cells. A disease in which the body cannot control the amount of glucose (a type of sugar) in the blood, causing the kidneys to produce large amounts of urine. This disease occurs when the body does not produce enough insulin or does not use it properly. The most common forms of diabetes are type 1 diabetes (5%), an autoimmune disease, and type 2 diabetes (95%) associated with obesity. Gestational diabetes is a form of diabetes that develops during pregnancy, while other forms of diabetes are very rare and are caused by a single gene mutation.

1.1 Types of Diabetes:

There are three main types of diabetes:-
1) Type1 diabetes
2) Type2 Diabetes
3) Gestational Diabetes (diabetes while pregnant).

1) Type1 diabetes:- If you have type1 diabetes, your body does not produce insulin. The immune system attacks and destroys the cells in the pancreas that produce insulin. Type 1 diabetes is usually diagnosed in children and young adults, but it can occur at any age. People with type1 diabetes must take daily insulin to survive. About 5–10% of people with diabetes have type1 diabetes.

2) Type2 diabetes:- When you have type 2 diabetes, your body cannot produce or use insulin well. Type 2 diabetes can develop at any age, even in childhood. However, this type of diabetes is most common among middle-aged and older people. Type2 is the most common type of diabetes. About 90–95% of people with diabetes have type2 diabetes. Type2 diabetes can be prevented or delayed through healthy lifestyle changes, such as losing weight, eating healthy food, and being active.

3) Gestational diabetes:- Gestational diabetes develops in some women when they are pregnant. Most of the time, this type of diabetes goes away after the baby is born. However, if you've had gestational diabetes, you have a greater chance of developing type2 diabetes later in life. Sometimes diabetes diagnosed during pregnancy is actually type2 diabetes.

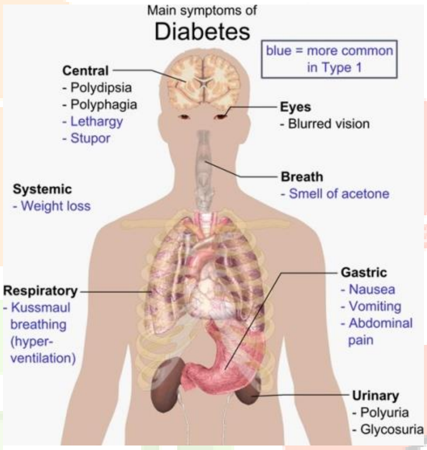
1.2 Pre-diabetes:

Pre- diabetes is a serious health condition in which blood sugar levels are higher than normal but not yet high enough to be diagnosed as type2 diabetes. About 96 million American adults (more than1in 3) have pre-diabetes. More than 80% of people with pre-diabetes don't know they have diabetes. Pre-diabetes increases your risk of type2 diabetes, heart disease, and stroke. The good news is that if you have prediabetes, the CDC's National Diabetes Prevention Program can help you prevent or delay type 2 diabetes and other serious health problems by making lifestyle changes.

Symptoms:- Being overweight, Being 45 years of age or older, Having a parent, brother, or sister with type2 diabetes Being physically active less than three times a week

Symptoms of Diabetes-

- Feeling more thirsty than usual.
- Urinating often.
- Losing weight without trying.
- Presence of ketones in the urine. Ketones are a byproduct of the breakdown of muscle and fat that happens when there's not enough available insulin.
- Feeling tired and weak.
- Feeling irritable or having other mood changes.
- Having blurry vision.
- Having slow-healing sores.
- Getting a lot of infections, such as gum, skin and vaginal infections.



1.3 Cause of Diabetes:

Too much glucose circulating in your bloodstream causes diabetes, regardless of the type. However, the reason why your blood glucose levels are high differs depending on the type of diabetes**. Causes of diabetes include:**

Insulin resistance: Type 2 diabetes is usually caused by insulin resistance. Insulin resistance occurs when cells in muscle, fat, and the liver do not respond to insulin as they should. Many factors and conditions can cause varying degrees of insulin resistance, including obesity, physical inactivity, diet, hormonal deficiencies, genetics, and certain medications.

Autoimmune disease: Type1 diabetes and LADA occur when your immune system attacks the insulin-producing cells in your pancreas.

Hormonal imbalances: During pregnancy, the placenta releases hormones that cause insulin resistance. If your pancreas does not produce enough insulin to overcome insulin resistance, you can develop diabetes. Other hormone-related disorders, such as acromegaly and Cushing's syndrome, can cause type2 diabetes.

Pancreatic damage: Physical damage to the pancreas due to disease, surgery, or injury can impair the pancreas's ability to produce insulin, leading to type3c diabetes.

Genetic mutations: Certain genetic mutations can cause MODY and neonatal diabetes.

2. DIABETES DETECTION WITH MACHINE LEARNING:

Machine learning is a technique by which computing systems learn the characteristics of input data. These methods have proven effective in detecting diabetes. Many machine learning algorithms have been developed, including supervised, unsupervised, and reinforcement learning methods. Since machine learning methods are data-driven, this is obviously practical. With huge amounts of data being loaded into databases, machine learning can save a significant amount of human effort. The model is trained on this data and provides the most appropriate result based on the input data. Models can be trained on any parameter that is acceptable in terms of practicality and medical requirements. Some of them examine facial features; others look for blood history data obtained from patients. Since there are many symptoms of the disease, the parameters vary accordingly. Using the various proposed methods, the researchers explored different algorithms and tuned numerous shy per parameters to obtain results that seemed most suitable for real-world applications.

2.1 Techniques Used For Diabetes Detection:

1) K-Nearest Neighbor

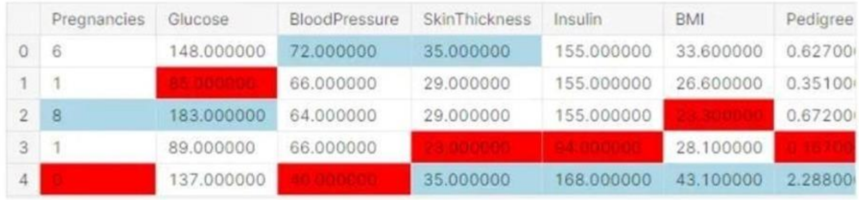
Description of Dataset:
I obtained the data from Kaggle, but the National Institute of Diabetes and Digestive and Kidney Diseases was the source of the information. This data covers several indicators and findings that describe whether or not a person has diabetes. This paper contains a summary of research from various patients that determined if a patient had diabetes or not. In this course, I'll examine some of the information provided to patients to check if they have diabetes utilizing this data and the KNN algorithm. We will manage, reject, and clean up the 768 studies in this database with diabetic and non diabetic individuals in order to use the minor KNN prediction model.

A supervised machine learning approach that deals with similarity is **the KNN algorithm**. KNN stands for K's closest neighbor. It is a classification algorithm that determines the class of thetarget variable using the nearest neighbor's numerical definition. It determines the distancebetween each sample of the training data and the sample you want to distribute, and then it categorizes your sample into the k classes that are closest in terms of the most prevalent 6 classes.

We use Euclidean distance to measure the distance between two data points or vectors from the dataset.

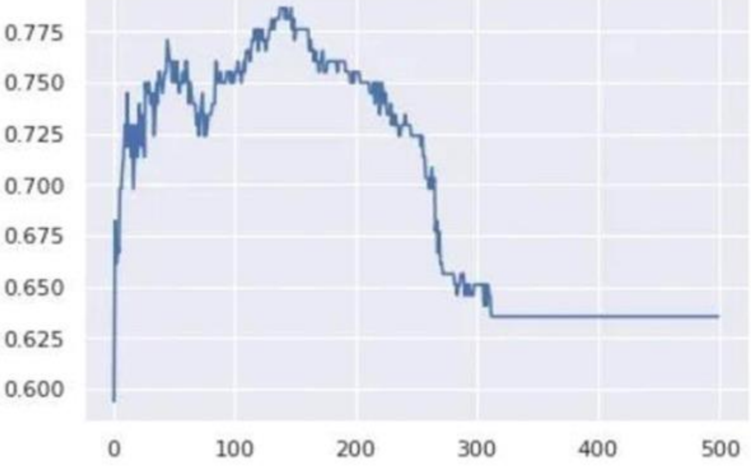
$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2}$$

Manipulating and Cleaning our dataset: Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Pedigree, are the most important data with a visible impact which determine if a patient is diabetic or not.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	Pedigree
0	6	148.000000	72.000000	35.000000	155.000000	33.600000	0.62700
1	1	85.000000	66.000000	29.000000	155.000000	26.600000	0.35100
2	8	183.000000	64.000000	29.000000	155.000000	23.300000	0.67200
3	1	89.000000	66.000000	23.000000	94.000000	28.100000	0.16700
4	0	137.000000	40.000000	35.000000	168.000000	43.100000	2.28800

I will test and plot the model with K values from 1 up to 500 and see where are we with the best overall k values



Having to experiment with different K from n=1 to n=500 , From the figure I can conclude that the best k that could optimize this model is between 100 to 200 offering a 77% accuracy . The ideal k value for this dataset should be 120 give or take.

2) Support Vector Machine (SVM)

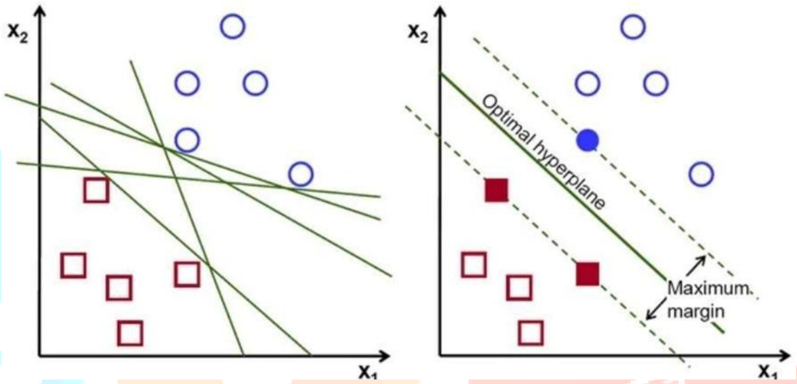
Preparing Our Training Data:

The Pima Indian Diabetes database will serve as the training set for this issue. Many of the causes and effects of diabetes are discussed in this document. The outcome shows whether the subject has diabetes (1) or not (0). These forecasts are known as features in machine learning.

Support Vector Machine Algorithm:

One of the most well-liked supervised learning methods for classification and regression issues is called Support Vector Machines, or SVMs. However, classification issues in machine learning frequently employ it. The SVM algorithm's goal is to draw a solid line or decision boundary that can divide the n-dimensional space into classes, allowing us to accurately and conveniently add additional data in the future. The hyperplane is the name of this well defined boundary. SVM chooses point clouds and/or vectors to build a hyperplane.

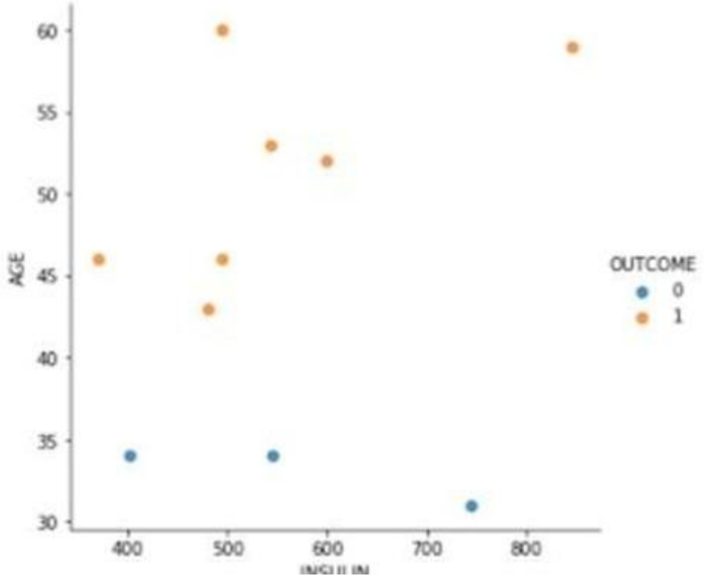
The algorithm is known as a vector machine because of these criteria, which are referred to as support vectors. Take a look at the illustration below, which shows two distinct units separated into general or decision planes.



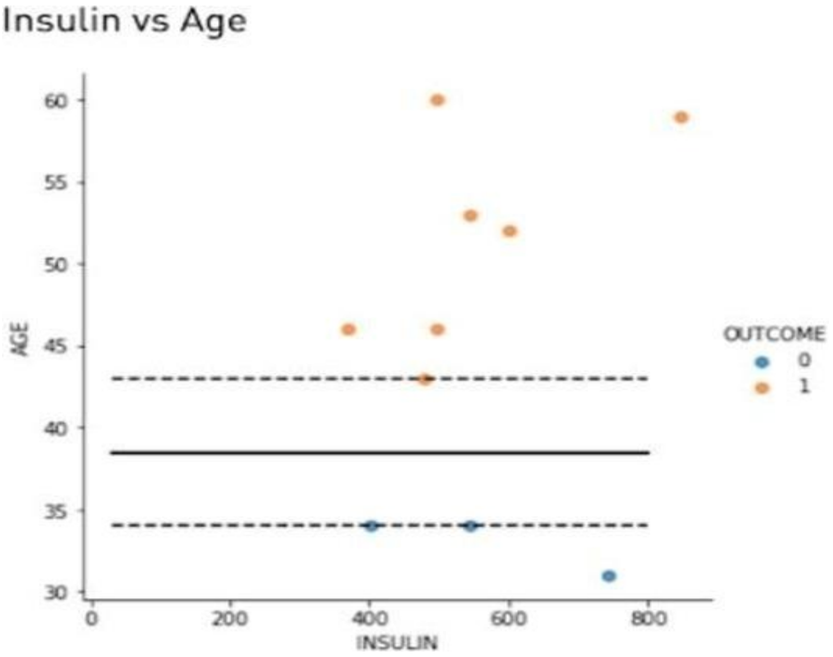
Possible hyperplanes

Visualizing the Hyperplane and Support Vectors:

Let's use just 2 dimensions to visualize our hyperplane: insulin levels and age, since we are unable to visualize data when there are so many dimensions. To get a separation without misclassifications for illustrative reasons, we can filter the data to only include individuals who are atleast 30 years old and have serum insulin levels exceeding 350µU/ml.



Results in the below hyperplane and parallels (dotted lines) passing through the Support Vector.



3) Decision Tree:

A supervised learning method called a decision tree can be used to solve classification and regression problems, but it is typically favored for doing so. It is a graphical depiction for obtaining all feasible answers to a choice or problem based on predetermined conditions. The CART algorithm, which stands for Classification and Regression Tree algorithm, is used toconstructa tree.

Algorithm:

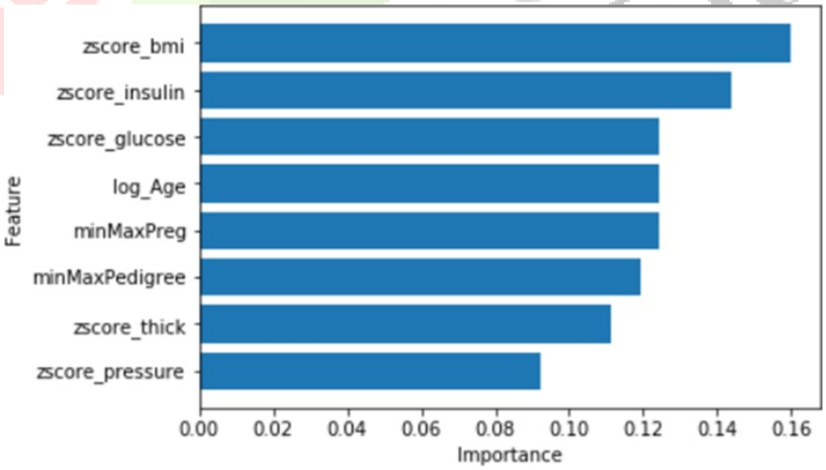
Step-1: Begin the tree with the root node, says S, which contains the complete dataset.

Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).

Step-3: Divide the S into subsets that contains possible values for the best attributes.

Step-4: Generate the decision tree node, which contains the best attribute.

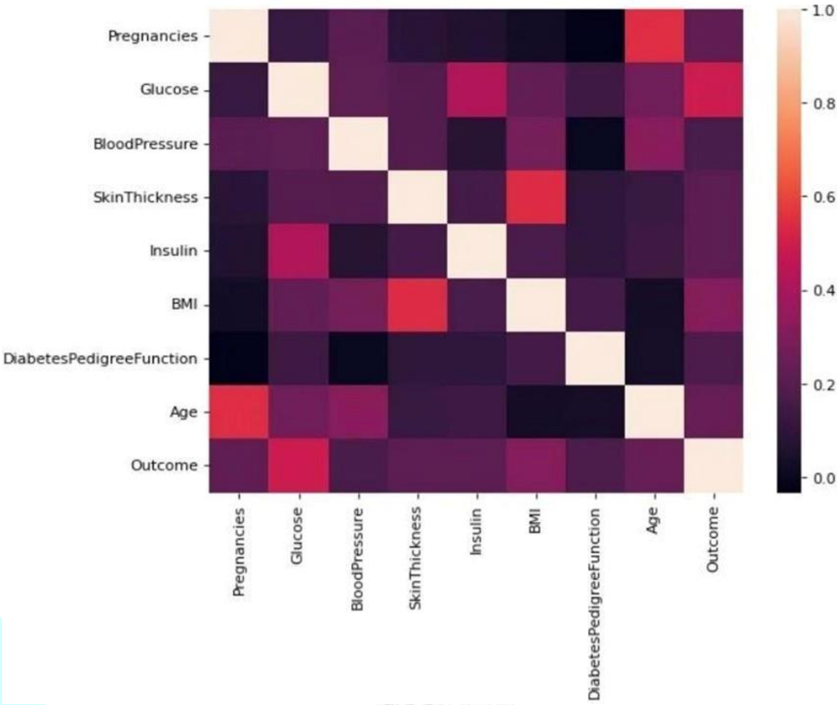
Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

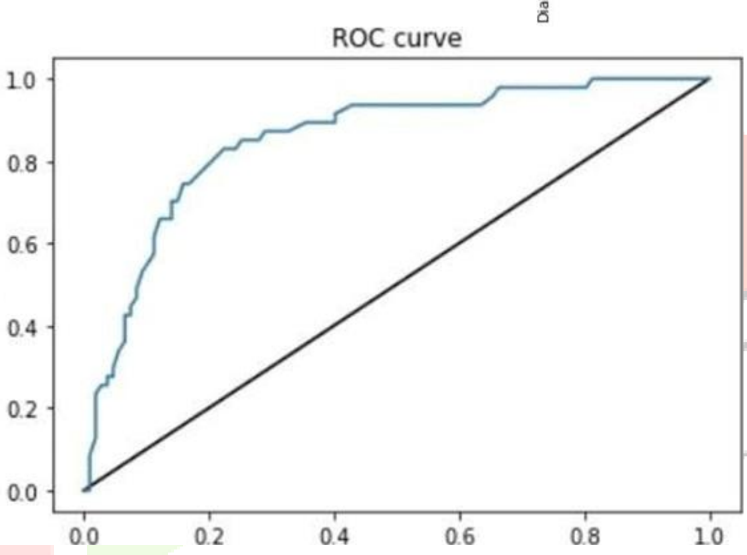


4) RandomForest:-

Instead of relying on one decision tree, the random forest takes the prediction from each tree and bases its prediction of the final output on the majority votes of predictions. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average toimprove the predictive accuracy of thatdataset.

Data Visualization: The correlation between each columns are visualized using heatmap.From the output, the lighter colors indicate more correlation. We notice the 10 correlation between pairs of features, like age and pregnancies, or BMI and skin thickness, etc.





For our model, the Area Under the Receiver Operating Characteristic Curve (ROCAUC) scoreis85%. This implies that the classification model is good enough to detect the diabetic patient.

3. CONCLUSION:

The main purpose of this project is to develop and implement diabetes prediction using machine learning and evaluate the performance of this successfully completed model. The proposedmethod uses various classification and learning methods using SVM, KNN, random forests, anddecision trees. And it achieved a classification accuracy of 77%. Test results can help patients make early predictions and decisions to treat diabetes and save lives.

In conclusion, machine learning algorithms have the potential to help healthcare practitioners identify people who may be at risk of acquiring diabetes or who have undiagnosed diabetes. These algorithms understand patterns and trends that may be suggestive of diabetes and can forecast a person's chance of contracting the condition by analyzing vast volumes of patient data.

The research and evaluations on machine learning-based diabetes detection systems have yielded encouraging results, with high accuracy rates reported in numerous studies. To validate these findings on larger datasets and various populations, additional study is required.

4. REFERENCES:

[1]. S. Arora, S. Karthikeyan, & I. Bose (2019). Using electronic health information, a machine learning-based technique is used to diagnose diabetes. 98, 103277, Journal of Biomedical Informatics.

[2].https://www.google.com/url?sa=t&source=web&rct=j&url=https://www.analyticsvidhya.com/blog/2022/01/diabetes-prediction-using-machinelearning/&ved=2ahUKEwi514GhqszAhXRbWwGHUbsAicQFnoECAoQAQ&usg=AOvVaw148KuG1fv9gVB4bLSyUjBU

[3]. Gupta, P. Kaur, and M. Kaur (2020). Review of machine learning for diabetes prediction. Medical artificial intelligence, 102, 101753. Healthcare Engineering.

[4]. https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes

[5].https://www.google.com/url?sa=t&source=web&rct=j&url=https://medium.com/codex/dia betes-predication-system-with-knn-algorithme040999229f7&ved=2ahUKEwipq7C3qszAhUOUGwGHfYoAaMQFnoECAwQAQ&usg=AOvVaw2v0eqgh4QVkpSYSPt5t3s

[6]. Patel, V., Khande, R., and Narkhede, B. machine learning algorithm for diabetes prediction. 8(2):55–60 International Journal of Computer Science and Mobile Computing