IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁙ Impact Factor 8.102 ⁙ Peer-reviewed / Refereed journal ⁙ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12725

Early Detection of Diabetes Using Machine Learning

Gururaj Mannolkar[1], Prof. Hrishikesh Mogare[2]

Student, MCA Department, KLS GIT, Belgaum, India[1]

Associate Professor (Mentor), MCA Department, KLS GIT, Belgaum, India[2]

Abstract: Diabetes is a long-term health condition that affects many people worldwide. It results in high levels of sugar in the blood, leading to symptoms like frequent urination, increased thirst, and hunger. If not managed properly, diabetes can cause serious complications like blindness, kidney failure, amputations, heart problems, and stroke.

Our body normally uses a hormone called insulin to regulate blood sugar levels. However, in diabetes, this system malfunctions. There are mainly two types of diabetes: type 1 and type 2. Type 1 diabetes occurs when the body does not produce enough insulin, while type 2 diabetes happens when the body does not use insulin effectively. Additionally, there is gestational diabetes, which develops during pregnancy. To tackle the growing concern of diabetes, researchers are exploring ways to use machine learning, to predict diabetes at an early stage with high accuracy. Machine learning helps machines learn from past data and experiences, and it can be useful in analyzing complex patterns in large datasets.

The goal of this project is to create a system that can predict diabetes in patients with better accuracy by combining the results of various machine-learning techniques. These techniques include K nearest neighbor, Logistic Regression, Random Forest, Support Vector Machine, and Decision Tree algorithms. Each algorithm is applied to the data, and its accuracy in predicting diabetes is calculated. The algorithm that shows good accuracy in its predictions will be chosen as the model for predicting diabetes in patients.

By using machine learning to predict diabetes early on, doctors and healthcare professionals can intervene timely and provide appropriate treatment to manage the condition effectively. This can help prevent or reduce the severity of complications associated with diabetes, thus improving the overall health and quality of life for affected individuals.

In summary, the project aims to utilize machine learning to develop an accurate and reliable system for early diabetes prediction, ultimately contributing to better healthcare outcomes for individuals at risk of diabetes.

Keywords: Diabetes, Blood sugar, Insulin, Type 1 diabetes, Type 2 diabetes, Gestational diabetes, Machine learning, Predictive modeling, K nearest neighbor, Logistic Regression, Random Forest, Support Vector Machine, Decision Tree, Early prediction, Healthcare intervention, Complications, Quality of life, Healthcare outcomes, Timely treatment, Patient risk assessment.

I. INTRODUCTION

Diabetes is a rapidly growing disease affecting people of all age groups, including young individuals. To understand diabetes, we must first grasp how our bodies function without the condition. When we eat foods containing carbohydrates, like bread, pasta, rice, fruits, and starchy vegetables, they get broken down into glucose (sugar), which serves as the body's main energy source.

This glucose travels through our bloodstream, some of it reaching our brain to help us think and function, while the rest is taken to the body's cells for energy or stored in the liver for later use. To use glucose for energy, our body needs a hormone called insulin, which acts like a key to unlock the cells' doors. Insulin is produced by the pancreas' beta cells.

However, in diabetes, things don't work smoothly. There are two main problems that can lead to diabetes. First, the pancreas may not produce enough insulin (insulin deficiency). Second, even if the pancreas produces enough insulin, the body may not respond to it properly (insulin resistance). In both cases, glucose starts to accumulate in the bloodstream, leading to high blood sugar levels, which is called hyperglycemia. This is when diabetes develops. Diabetes Mellitus refers to the condition where there are high levels of sugar (glucose) in the blood and urine.

DOI: 10.17148/IJARCCE.2023.12725

In summary, diabetes occurs when our body faces difficulties in regulating blood sugar levels due to a lack of insulin or the body's inability to use insulin effectively. It's essential to understand these processes to comprehend how diabetes develops and how it affects our health.

Types of Diabetes :

- Type 1 Diabetes: In Type 1 diabetes, the immune system mistakenly attacks and damages the cells in the pancreas responsible for producing insulin. As a result, the body cannot produce enough insulin to regulate blood sugar levels properly. This type of diabetes is usually diagnosed in children and young adults. The exact cause of Type 1 diabetes is not fully understood, and there are currently no known ways to prevent it.

- Type 2 Diabetes: Type 2 diabetes is the most common form of diabetes, accounting for around 90% of all diabetes cases. In Type 2 diabetes, the body either doesn't produce enough insulin or the cells become resistant to insulin's actions. This means that even though insulin is present, it's not effective in allowing glucose to enter the cells for energy. Type 2 diabetes is often associated with genetic factors, such as a family history of diabetes, and lifestyle factors, including poor diet and lack of physical activity.

- Gestational Diabetes: Gestational diabetes occurs in pregnant women who develop high blood sugar levels during pregnancy. This condition typically arises in the second or third trimester and usually goes away after giving birth. However, women who have had gestational diabetes are at an increased risk of developing Type 2 diabetes later in life. Additionally, there's a higher chance of gestational diabetes recurring in future pregnancies.

In summary, Type 1 diabetes results from the immune system attacking insulin-producing cells, Type 2 diabetes arises due to insufficient insulin production or insulin resistance, and gestational diabetes affects pregnant women with high blood sugar levels during pregnancy, potentially leading to Type 2 diabetes later in life. Understanding these types of diabetes helps in managing and preventing complications associated with the condition.

Symptoms of Diabetes:

- Frequent Urination: People with diabetes may experience the need to urinate more often than usual. This happens because the high levels of sugar in the blood can spill into the urine, causing the body to try to get rid of the excess sugar through urine.

- Increased Thirst: Excessive thirst is a common symptom of diabetes. When blood sugar levels are high, the body tries to compensate by extracting more fluids from the body tissues, leading to dehydration and increased thirst.

- Tiredness/Sleepiness: Diabetes can cause feelings of tiredness and fatigue due to the body's inability to effectively use glucose for energy, leading to an energy deficit.

- Weight Loss: Unexplained weight loss can occur in some individuals with diabetes. This happens because the body breaks down muscle and fat for energy since it can't properly use glucose.

- Blurred Vision: High blood sugar levels can affect the lens in the eye, leading to temporary changes in vision, causing blurriness.

- Mood Swings: Diabetes can sometimes affect mood, causing irritability, mood swings, and even depression due to hormonal imbalances caused by fluctuating blood sugar levels.

- Confusion and Difficulty Concentrating: Rapid changes in blood sugar levels can impact cognitive function, leading to confusion and difficulty concentrating.

- Frequent Infections: High blood sugar weakens the immune system, making individuals with diabetes more susceptible to infections. Common infections include urinary tract infections, skin infections, and yeast infections.

It's important to note that not everyone with diabetes will experience all of these symptoms, and some symptoms may vary in severity depending on individual factors. If you notice any of these symptoms, it's essential to seek medical attention for proper evaluation and diagnosis. Early detection and management of diabetes are crucial for preventing complications and maintaining overall health.

DOI: 10.17148/IJARCCE.2023.12725

Causes of Diabetes:

- Genetic Factors: Diabetes can be influenced by genetic factors, which means it may run in families. Certain genes on chromosome 6 play a significant role in how the body responds to various antigens. If there are specific mutations in these genes, it can lead to an increased risk of developing diabetes. This is particularly true for Type 1 diabetes, where genetic factors are a key contributing factor.

- Viral Infections: Some viral infections have been associated with an increased risk of developing both Type 1 and Type 2 diabetes. Studies have shown that infections with viruses like rubella, Coxsackievirus, mumps, hepatitis B virus, and cytomegalovirus can influence the development of diabetes. These infections can trigger an autoimmune response in the body, where the immune system mistakenly attacks and damages the insulin- producing cells in the pancreas, leading to Type 1 diabetes. Additionally, viral infections can also cause inflammation and insulin resistance, contributing to the development of Type 2 diabetes.

It's important to note that while genetic factors and viral infections play a role in the development of diabetes, they are not the only factors. Lifestyle factors, such as diet, physical activity, and body weight, also significantly influence the risk of developing Type 2 diabetes. For individuals with a family history of diabetes or those who have had viral infections, adopting a healthy lifestyle can help reduce the risk and manage the condition effectively. Regular medical checkups and early detection are essential for timely intervention and diabetes management.

II. LITERATURE REVIEW

1. Yasodha and their team conducted a study to classify whether a person has diabetes or not using different types of datasets. They collected data from a hospital warehouse, which included information from 200 individuals with diabetes and non-diabetic individuals. The data consisted of nine attributes related to blood and urine tests. To analyze the data, they used a tool called WEKA, which helps in classifying and evaluating data. They employed a technique called "10-fold cross-validation," which is a way to assess the model's performance using smaller datasets. They tested four different algorithms (methods) for classification: Naïve Bayes, J48, REP Tree, and Random Tree. After comparing the results, they found that the J48 algorithm performed the best, achieving an accuracy of 60.2% compared to the other methods.

2. Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm
Authors: Samrat Kumar Dey, Ashraf Hossain, and Md. Mahbubur Rahman
Published In: 2018 21st International Conference of Computer and Information Technology (ICCIT)

3. Tshepo Sr and their team used a decision tree algorithm called CART on a diabetes dataset. Before applying the algorithm, they addressed the issue of class imbalance in the data to improve accuracy rates. Class imbalance occurs when a dataset has two possible outcomes for the class variable, but one outcome is much more frequent than the other. This can affect the performance of predictive models. To handle this problem, Tshepo Sr and their team applied a resample filter to balance the classes in the dataset during data preprocessing. By addressing the class imbalance issue early in the process, they were able to boost the accuracy of the predictive model when using the decision tree algorithm on the diabetes dataset.

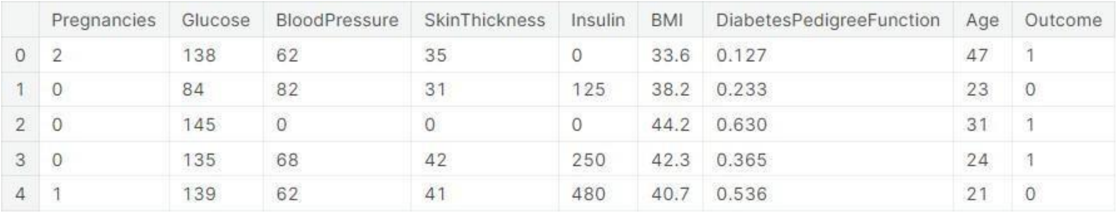
III. METHODOLOGY

In this section, we will discuss the methods used in machine learning to predict diabetes and how we propose to improve accuracy. The study uses five different methods, which are explained below. The goal is to measure the accuracy of these machine learning models, and once the model is created, it can be used for making predictions.

Dataset Description :
The dataset used in this study was obtained from the website "https://www.kaggle.com/johndasilva/diabetes." It includes information from 2000 cases related to diabetes. The main goal of the study is to use this data to predict whether a patient is diabetic or not based on certain measures or features.

1. Pregnancies: Number of times the patient has been pregnant.
2. Glucose: Blood glucose level of the patient.
3. BloodPressure: Blood pressure measurement of the patient.
4. SkinThickness: Thickness of the skinfold at the triceps area.
5. Insulin: Insulin level in the patient's body.

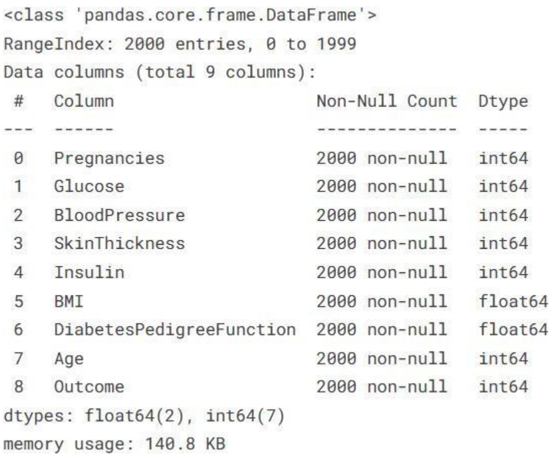
6. BMI (Body Mass Index): A measure of body fat based on weight and height.
7. DiabetesPedigreeFunction: A function that represents



	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 features in the dataset the likelihood of diabetes based on family history.

1. Age: Age of the patient.
2. Outcome: This indicates if the patient has diabetes (1) or does not have diabetes (0).



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               2000 non-null   int64  
 1   Glucose                   2000 non-null   int64  
 2   BloodPressure             2000 non-null   int64  
 3   SkinThickness             2000 non-null   int64  
 4   Insulin                   2000 non-null   int64  
 5   BMI                       2000 non-null   float64
 6   DiabetesPedigreeFunction  2000 non-null   float64
 7   Age                       2000 non-null   int64  
 8   Outcome                   2000 non-null   int64  
dtypes: float64(2), int64(7)
memory usage: 140.8 KB
```

Fig. 2 This dataset has no null values

We can observe that none of the features have a very strong correlation with the outcome value. Some features show a negative correlation, meaning they tend to decrease when the outcome value increases, while others show a positive correlation, meaning they tend to increase when the outcome value increases. However, none of the features alone can decisively predict whether a person has diabetes or not.

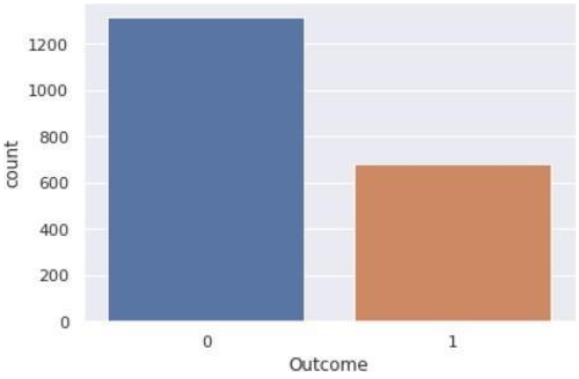


Fig. 3 Count VS Outcome

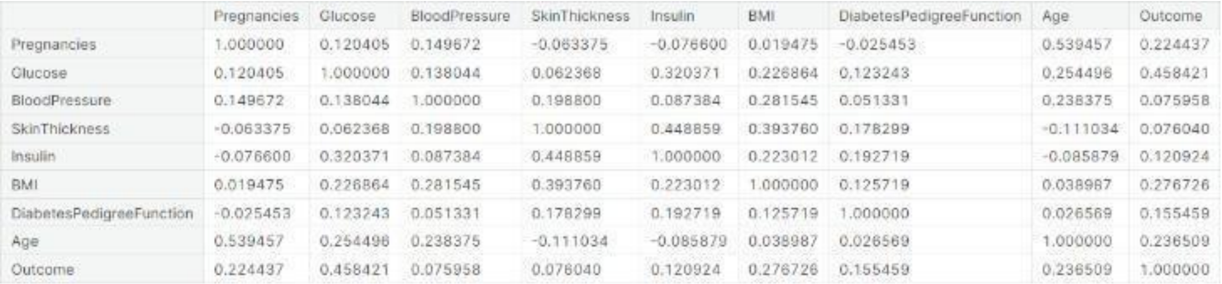
The graph indicates that the data is imbalanced, with more data points showing an outcome value of 0, which means diabetes was not present. The number of non-diabetic cases is almost twice as many as the number of diabetic cases in the dataset.

DOI: 10.17148/IJARCCE.2023.12725

k-Nearest Neighbors:

The k-NN algorithm is one of the simplest machine learning methods. To create the model, it just needs to store the training data set. When we want to make a prediction for a new data point, the algorithm looks for the closest data points (nearest neighbors) in the training data set to the new data point. The prediction is then based on the majority class of these nearest neighbors. In simple terms, k-NN compares the new data point with known data points and predicts its class based on the classes of the closest data points it finds.

Fig. 4 k-Nearest Neighbors



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.120405	0.149672	-0.063375	-0.076600	0.019475	-0.025453	0.539457	0.224437
Glucose	0.120405	1.000000	0.138044	0.062368	0.320371	0.226864	0.123243	0.254496	0.458421
BloodPressure	0.149672	0.138044	1.000000	0.198800	0.087384	0.281545	0.051331	0.238375	0.075958
SkinThickness	-0.063375	0.062368	0.198800	1.000000	0.448859	0.393760	0.178299	-0.111034	0.076040
Insulin	-0.076600	0.320371	0.087384	0.448859	1.000000	0.223012	0.192719	-0.085879	0.120924
BMI	0.019475	0.226864	0.281545	0.393760	0.223012	1.000000	0.125719	0.038987	0.276726
DiabetesPedigreeFunction	-0.025453	0.123243	0.051331	0.178299	0.192719	0.125719	1.000000	0.026569	0.155459
Age	0.539457	0.254496	0.238375	-0.111034	-0.085879	0.038987	0.026569	1.000000	0.236509
Outcome	0.224437	0.458421	0.075958	0.076040	0.120924	0.276726	0.155459	0.236509	1.000000

Fig. 5 k-Nearest Neighbors

In the plot above, we can see the relationship between the complexity of the model (controlled by the parameter "n_neighbors") and its accuracy. The x-axis represents the number of neighbors considered, while the y-axis shows the training and test set accuracy.

When we use only one single nearest neighbor, the model perfectly predicts the training data. However, this high accuracy on the training data may not necessarily mean the model is the best. As we increase the number of neighbors considered, the training accuracy drops. This suggests that relying on just one neighbor leads to a model that is too complex, and it may not generalize well to new, unseen data.

The best performance is achieved around 9 neighbors, where both the training and test set accuracies are relatively high.In this case, the model is not too complex, and it can make accurate predictions on new data points.In simple terms, the plot shows how the number of neighbors affects the accuracy of the k-NN model. Using only one neighbor leads to a model that is too complex and overfits the training data. As we increase the number of neighbors, the model becomes more balanced and achieves the best performance around 9 neighbors, where both training and test accuracies are relatively high (around 0.81 for training and 0.78 for testing). This means the model is accurate and can generalize well to new data.

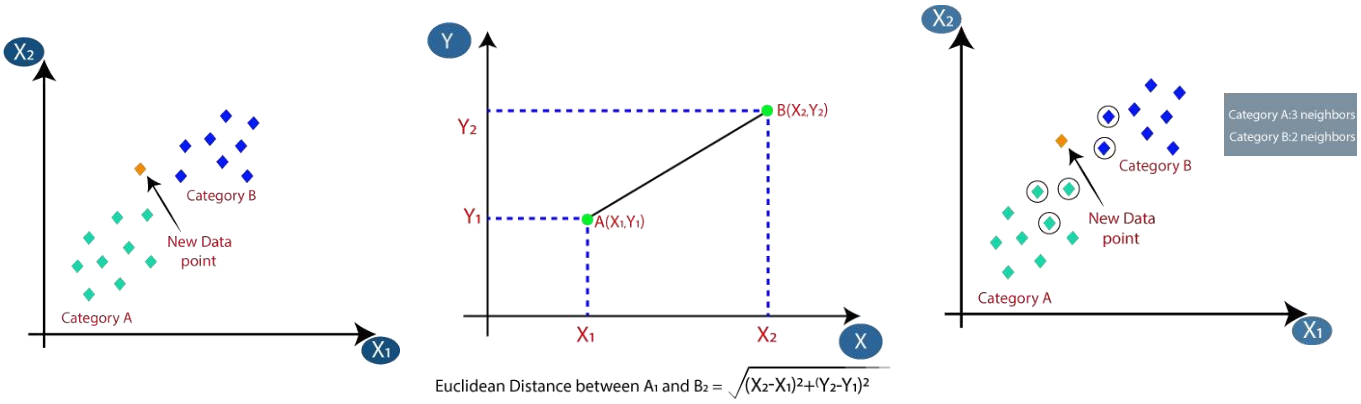


Fig. 6 k-Nearest Neighbors working

Training Accuracy	0.8
Testing Accuracy	0.7

TABLE 1 ACCURACY SCORE

Logistic regression:
Linear regression is a simple and popular machine learning algorithm used for predictive analysis. It is used to make predictions for continuous or numeric variables like sales, salary, age, and product price.

In linear regression, we look for a straight line that best represents the relationship between a dependent variable (the one we want to predict) and one or more independent variables (the ones we use to make predictions). The model finds how the value of the dependent variable changes based on the values of the independent variables.

The linear regression model provides a straight line on a graph that represents this relationship between the variables. It helps us predict the value of the dependent variable based on the independent variables.

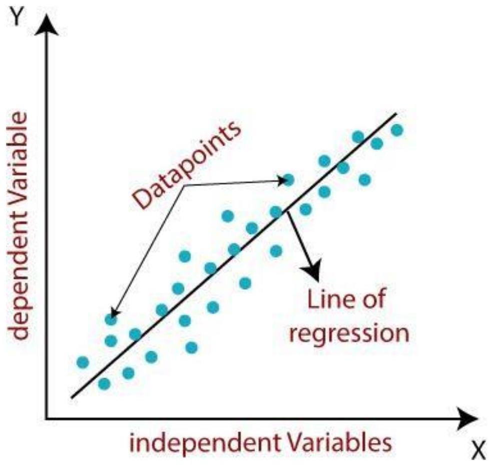


Fig. 7 Linear regression Mathematically, the linear regression equation looks like this: $Y = a0 + a1x + \varepsilon$

DOI: 10.17148/IJARCCE.2023.12725

Where:

- Y is the dependent variable (the one we want to predict).
- X is the independent variable (the one used for predictions).
- a0 is the intercept of the line (gives an additional degree of freedom to the line).
- a1 is the linear regression coefficient (a scale factor to each input value).
- ε is the random error that accounts for any uncertainty in the data.

The values of x and y in the equation are the training datasets used to represent the linear regression model. In simpler terms, linear regression helps us find a straight line that best describes the relationship between variables so we can make accurate predictions. The table shows the training and testing accuracies of the model for different values of the regularization parameter, denoted by C.

In the first row, when C is set to its default value of 1, the model achieves 77% accuracy on the training data and 78% accuracy on the testing data. In the second row, using C=0.01, the model achieves 78% accuracy on both the training and testing sets. When C is set to 100 in the third row, the training accuracy is a bit lower at 77.8%, while the testing accuracy is slightly higher at 79.2%.

The table suggests that the default value of C=1 provides reasonably good accuracy on both the training and testing data. Changing the value of C doesn't result in significant improvements, indicating that a more complex model with less regularization may not necessarily generalize better than the default setting.

In simple terms, Logistic Regression is a widely used algorithm for classification tasks. The table shows how the model performs with different values of the regularization parameter C. After comparing the results, it appears that the default value of C=1 is the best choice as it provides good accuracy on both the training and testing data. Changing C doesn't make a significant difference in the model's performance. Therefore, it's better to stick with the default value of C=1.

	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 ACCURACY SCORE

Decision Tree:

A Decision Tree is a supervised learning technique used for solving classification and regression problems. It is mainly preferred for classification tasks. The Decision Tree is like a tree with branches representing decision rules and leaves representing outcomes.

In a Decision Tree, there are two types of nodes: Decision Nodes and Leaf Nodes. Decision Nodes are used to make decisions based on the features of the data, and they have multiple branches. On the other hand, Leaf Nodes are the final outcomes without any further branches.

The Decision Tree makes decisions or tests based on the features of the data. It represents all the possible solutions to a problem based on given conditions.

It starts with a root node and expands into branches, constructing a tree-like structure. The process of building the tree is done using the CART algorithm, which stands for Classification and Regression Tree algorithm.

The tree asks questions based on the data, and depending on the answers (Yes or No), it continues to split into subtrees. This process goes on until the tree reaches the leaves, which provide the final outcomes.

In simpler terms, a Decision Tree is like a tree that asks questions about the data and finds solutions step by step until it reaches the final answer. It is a useful tool for classifying data into different categories based on their features.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁑ Impact Factor 8.102 ⁑ Peer-reviewed / Refereed journal ⁑ Vol. 12, Issue 7, July 2023

DOI: 10.17148/IJARCCE.2023.12725

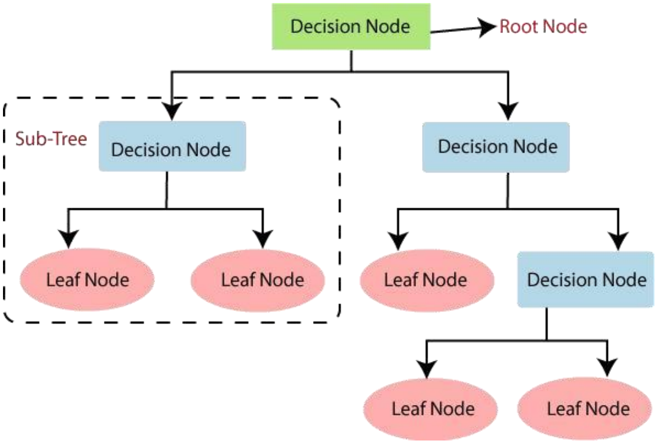


Fig. 8 Decision Tree

We can adjust the maximum number of features considered while building the model.

Training Accuracy	1.00
Testing Accuracy	0.9

TABLE 3 ACCURACY SCORE

In Table-3, it shows that the model achieved 100% accuracy on the training set and a high accuracy of 99% on the test set. This means the model is performing well and can make accurate predictions. In Decision Trees, we can also calculate feature importance. Feature importance rates how important each feature is in making decisions within the tree. It assigns a number between 0 and 1 to each feature, where 0 means the feature is "not used at all" in the decision-making, and 1 means the feature "perfectly predicts the target."

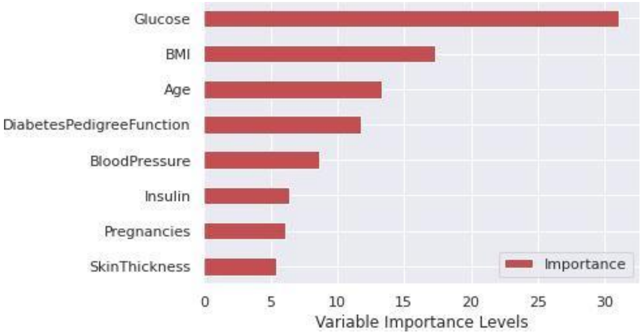


Fig. 9 Variable Importance Levels

In this case, the feature "Glucose" is found to be the most important feature, indicating that it plays a significant role in the decisions made by the Decision Tree model.

In simpler terms, Decision Tree is a classifier that builds a tree-like structure to predict the class values of data points. The model achieved a high accuracy of 99% on the test data, showing its effectiveness. Additionally, we can calculate feature importance to see which features are crucial for the model's decision-making. In this case, the feature "Glucose" is the most important, meaning it strongly influences the predictions made by the Decision Tree.

Random Forest:
Random Forest is a powerful classifier that builds a collection of decision trees, taking the concept of decision trees to the next level. In this method, a "forest" of trees is created, where each tree is formed by randomly selecting features from the total available features.

DOI: 10.17148/IJARCCE.2023.12725

In the study mentioned (Volume 6, Issue 4, May-June- 2020), the Random Forest achieved a training accuracy of 100%, meaning it perfectly predicted the training data. The testing accuracy was also high at 97.4%, indicating good performance on new, unseen data.

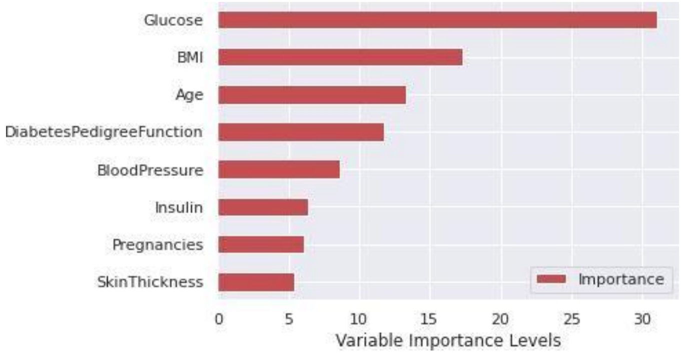


Fig. 10 Variable Importance Levels

When analyzing feature importance in the Random Forest, it was found that, like in a single decision tree, the "Glucose" feature was highly informative. However, the Random Forest also ranked the "BMI" feature as the second most important overall. In simple terms, Random Forest is an advanced version of decision trees. It creates a group of trees, each made with a random selection of features. The model achieved excellent accuracy in both training and testing, and it highlighted that "Glucose" was the most important feature for prediction. Additionally, the Random Forest considered "BMI" as the second most important feature in making predictions.

SVM:

Support Vector Machine (SVM) is a classifier that aims to create a hyperplane to effectively separate different classes of data by adjusting the distance between the data points and the hyperplane. The hyperplane is determined based on different kernels, such as linear, polynomial, radial basis function (rbf), and sigmoid.

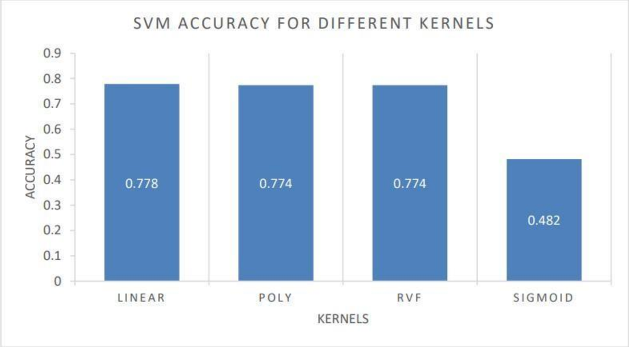


Fig. 11 SVM Accuracy for different kernels

the performance of SVM with different kernels was compared. The plot showed that the SVM with the linear kernel performed the best for the given dataset, achieving a score of 77% accuracy. In simple terms, SVM is a classifier that tries to find the best line or boundary (hyperplane) to separate data into different classes. The choice of the hyperplane depends on the kernel used. In this study, four different kernels were tested, and the plot showed that the linear kernel provided the best results with 77% accuracy on the dataset.

Accuracy Comparison:

Algorithms	Training Accuracy	Testing Accuracy
k-Nearest Neighbours	81	78
Logistic Regression	78	78
Decision Tree	98	99
Random Forest	94	97
SVM	76	77

TABLE 4 The table presents the accuracy values of five machine learning algorithms

DOI: 10.17148/IJARCCE.2023.12725

- k-Nearest Neighbors: It achieved 81% accuracy on the training data and 78% accuracy on the testing data.
- Logistic Regression: It obtained 78% accuracy on both the training and testing data.
- Decision Tree: This algorithm showed the highest accuracy, with 98% on the training data and 99% on the testing data.
- Random Forest: It achieved 94% accuracy on the training data and 97% accuracy on the testing data.
- SVM (Support Vector Machine): It obtained 76% accuracy on the training data and 77% accuracy on the testing data.

From the table, we can see that the Decision Tree algorithm performed the best with the highest accuracy. It achieved 98% accuracy on the training data and 99% accuracy on the testing data.

In simple terms, the table shows the accuracy of different machine learning algorithms on both training and testing data. The Decision Tree algorithm outperformed the others, with 98% accuracy on training data and 99% accuracy on testing data, making it the most accurate model for this particular dataset.

IV. CONCLUSION AND FUTURE WORK

In conclusion, this study focused on a crucial real-world medical problem - early detection of diabetes. The researchers designed a system for predicting diabetes using five different machine learning classification algorithms. They evaluated these algorithms on various measures using the john Diabetes Database. The experimental results showed that the system achieved an impressive accuracy of 99% using the Decision Tree algorithm.

In the future, this designed system and the machine learning algorithms used can be extended to predict or diagnose other diseases. The researchers suggest automating diabetes analysis using this system and exploring the potential of incorporating additional machine learning algorithms to further improve the accuracy and efficiency of disease prediction.

In simpler words, this study aimed to detect diabetes at an early stage. They developed a system that used five different machine learning methods to predict diabetes. The system achieved an accuracy of 99% using the Decision Tree algorithm. In the future, they plan to expand the system to detect other diseases and explore more machine learning techniques for improved disease prediction.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to **Prof. Hrishikesh Mogare** (Assistant Professor) from the Department of MCA, KLS Gogte Institute of Technology College for mentoring me.

I want to thank my mentor for inspiring and supporting me. Their active guidance, help, cooperation, and encouragement were instrumental in completing this work and writing the paper.

Lastly, I extend my heartfelt appreciation to my parents and family members who have always been a moral and financial support to me.

REFERENCES

[1] Improved J48 Classification Algorithm for the Prediction of Diabetes Authors: Gaganjot Kaur, Amit Chhabra
[2] Diagnosis of diabetes using classification mining techniques Author: Aiswarya Iyer, S. Jeyalatha, Ronak Sumbaly
[3] Diabetes prediction using Decision Tree
Author: Tshepo Sr.
[4] Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm
Authors: Samrat Kumar Dey, Ashraf Hossain, and Md. Mahbubur Rahman
[5] Diabetes Prediction Using Machine Learning Authors: KM Jyoti Rani