Fuzzy Systems and Soft Computing
ISSN : 1819-4362

IMPLEMENTING MACHINE LEARNING IN MEDICAL DIAGNOSIS: A SPECIAL REVIEW IN EARLY AGE DIABETES.

Dr. Monalisa Hati, Assistant Professor, Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Mumbai, Maharastra, India.
ssamit6@gmail.com.
Vaidehi Mahmia,Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Mumbai, Maharastra, India
Saunved Ghansham Giri, Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Mumbai, Maharastra, India
Rushabh Subhash Dorage, Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Mumbai, Maharastra, India

Abstract:

Diabetes is a chronic condition that has the potential to lead to a global health emergency. The International Diabetes Federation estimates that 382 million people worldwide have diabetes. This will quadruple 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. One of the main causes of blindness, renal failure, heart failure, stroke, and amputations is diabetes. Our bodies convert food into glucose, or sugars, when we consume. Our pancreas is then expected to release insulin. Insulin acts as a key to unlock our cells, letting glucose in and enabling us to Utilize the glucose to generate energy. However, this approach is ineffective in those with diabetes. The two most prevalent types of diabetes are type 1 and type 2, but there are other types as well, including gestational diabetes, which develops during pregnancy. The study of how machines learn from experience is the focus of the developing data science discipline of machine learning. The goal of this research 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 in the early stages. Decision trees, random forests, K closest neighbors, logistic regression, and support vector machines are among the methods that are employed. The model's accuracy when utilizing each of the algorithm is computed.

The model that predicts diabetes with the highest accuracy is then chosen.

Keywords: Machine Learning, Decision tree, K nearest neighbour, Logistic Regression, Support vector Machine

1.INTRODUCTION

Diabetes is a condition that is rapidly spreading, especially among young individuals. We must comprehend what occurs in the body in the absence of diabetes in order to comprehend the disease and how it develops. Glucose, or sugar, is derived from the meals we consume, particularly those high in carbohydrates. Everybody, including those with diabetes, requires carbohydrates since they are the body's primary energy source. Bread, cereal, pasta, rice, fruit, dairy products, and vegetables—particularly starchy vegetables—are examples of foods high in carbohydrates. After consuming these meals, the body converts them into glucose. The bloodstream carries the glucose throughout the body. A portion of the glucose is transported to the brain to support clear thinking and cognitive function. The rest of the glucose is transported to our body's cells for energy and to our liver, where it is stored for future use by the body. Insulin is necessary for the body to use glucose as fuel. The beta cells of the pancreas create the hormone insulin. Insulin functions similarly to a door key. To enable glucose to pass from the bloodstream via the cell's doors and into the cell, insulin binds itself to them. Diabetes occurs when the body is unable to use the insulin it generates (insulin resistance) or is unable to create enough of it (insulin deficit), causing glucose to accumulate in the circulation (hyperglycemia). High blood and urine sugar (glucose) levels are indicative of diabetes mellitus.

Causes of Diabetes

Diabetes is mostly caused by genetic causes. It is brought on by at least two mutated genes on chromosome 6, which controls how the body reacts to different antigens.

Type 1 and type 2 diabetes may also be influenced by viral infection. Rubella, Coxsackievirus, mumps, hepatitis B virus, and cytomegalovirus infections have been linked to an increased risk of diabetes, according to studies.

2.LITERATURE REVIEW

To determine if a person has diabetes or not, Yasodhaet al. [1] apply classification on a variety of dataset types. The data set for diabetic patients is created by compiling information from a hospital warehouse that has two hundred instances with nine different properties. Blood tests and urine tests are the two classes to which these instances of this dataset pertain. Because WEKA works so well on tiny datasets, it may be used in this study to categorize the data and evaluate the data using a 10-fold cross validation technique. The results are compared. The Random Tree, J48, REP Tree, and naïve Bayes are employed. J48 was found to be the most effective, with an accuracy of 60.2%.

In order to find ways to identify diabetes, Aiswarya et al. [2] use classification analysis with Decision Tree and Naïve Bayes algorithms to look at and analyze the patterns that appear in the data. The goal of the research is to provide a quicker and more effective way to diagnose the illness so that patients can receive timely treatment. with the PIMA dataset and a cross-validation methodology, the study found that the naïve Bayes algorithm achieves an accuracy rate of 79.5% with a 70:30 split, whereas the J48 algorithm achieves an accuracy rate of 74.8%.

In addition to attempting to evaluate and analyze the outcomes of many classification techniques in WEKA, Gupta et al. [3] seek to determine and compute the accuracy, sensitivity, and specificity % of various categorization methods. the effectiveness of the same classifiers using the same parameters (i.e., accuracy, sensitivity, and specificity) when applied to different programs, such as Rapidminer and Matlab. They used the BayesNet, JRIP, and Jgraft 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 Rapidminner.

After the data has been subjected to the resample filter. The author focuses on the issue of class imbalance and how it must be resolved before using any technique to increase accuracy rates. Class imbalance is typically seen in datasets with dichotomous values, meaning that the class variable has two alternative outcomes. This type of imbalance may be readily addressed if it is noticed early in the data preparation step and will increase the prediction model's accuracy.

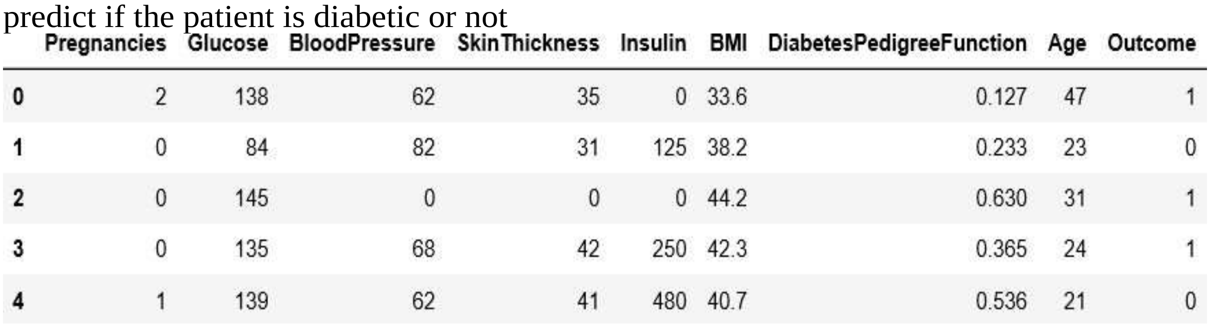
3.METHODOLOGY

The several classifiers used in machine learning to predict diabetes will be covered in this section. To increase the accuracy, we will also describe the methods we have suggested. This paper employed five different approaches. The definitions of the various techniques are given below. The machine learning models' accuracy metrics are the result. The model may then be applied to forecasting.

Dataset Description

The diabetes data set was originated from https://www.kaggle.com/johndasilva/diabetes.

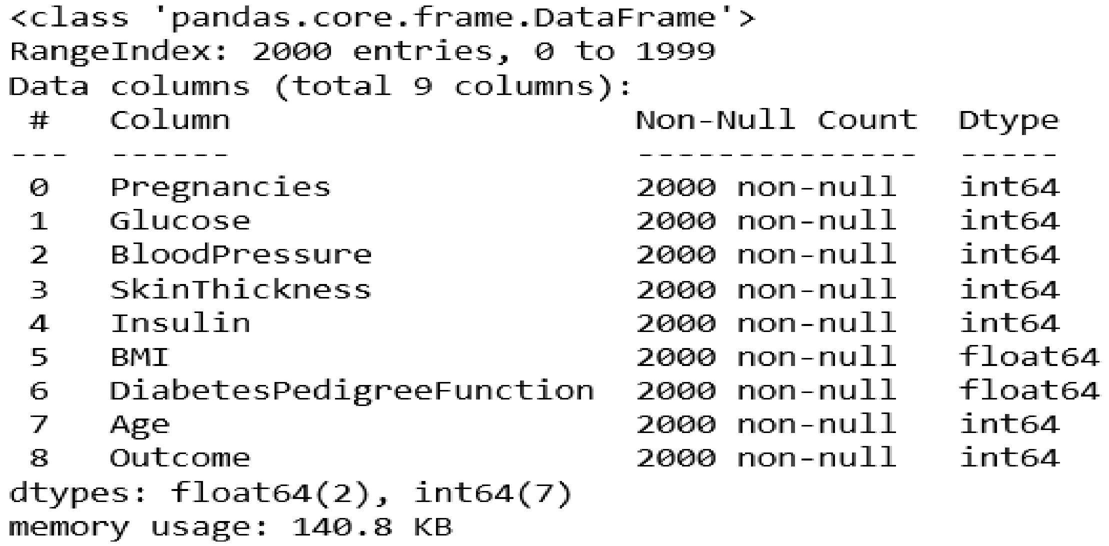
Diabetes dataset containing 2000 cases. The objective is to predict based on the measures to predict if the patient is diabetic or not



	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

-The diabetes data set consists of 2000 data points, with 9 features each.

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



```
<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
```

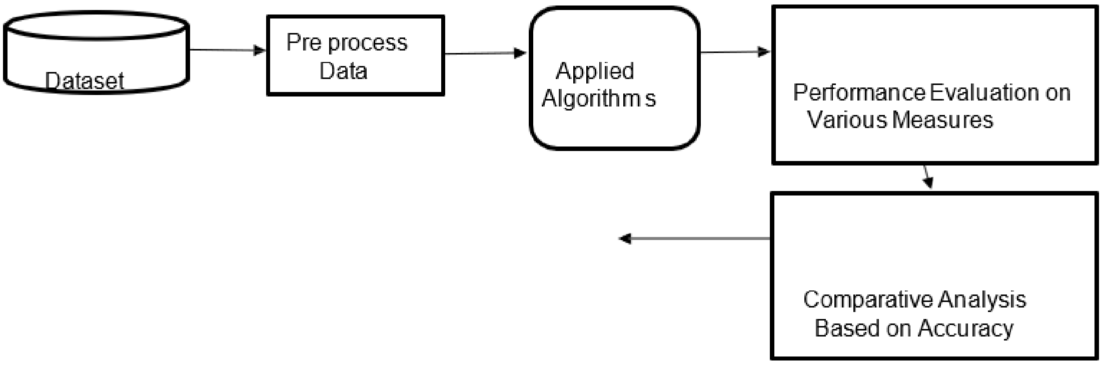
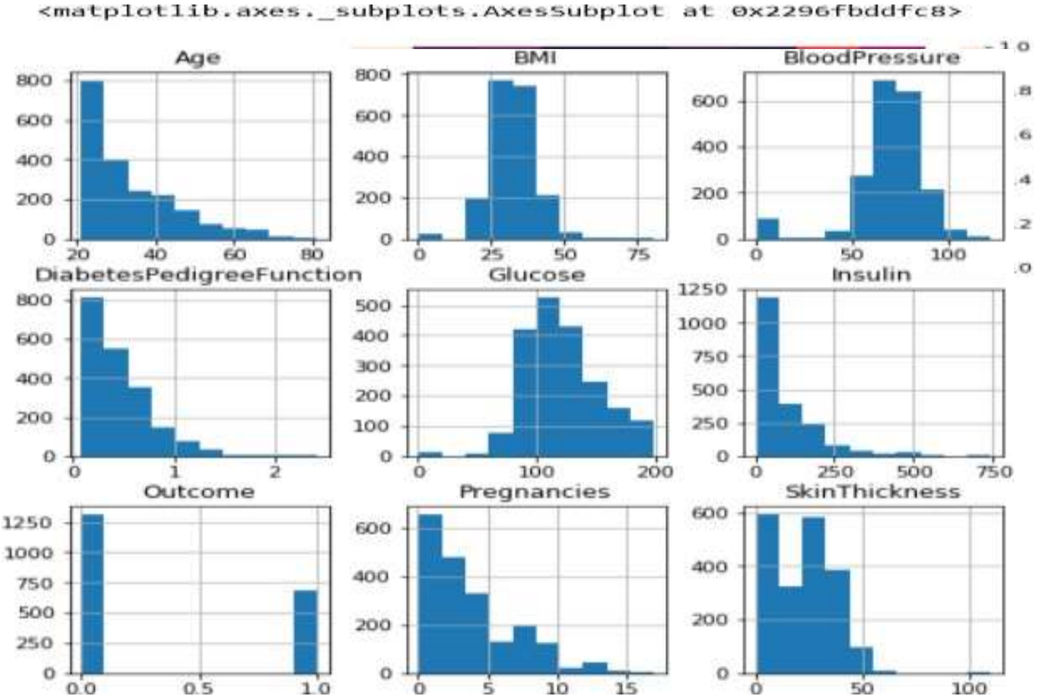

Dataset
Pre process Data
Applied Algorithm s
Performance Evaluation on Various Measures
Comparative Analysis Based on Accuracy

Fig 1: Proposed Model

Correlation Matrix:



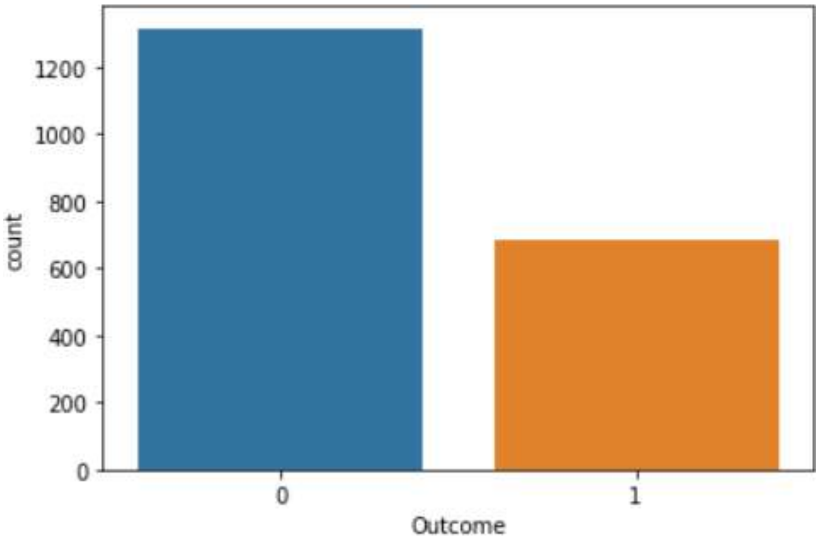
```
<matplotlib.axes._subplots.AxesSubplot at 0x2296fbddfc8>
```

The fact that no one attribute has a particularly strong association with our outcome value is readily apparent. The result value is negatively correlated with some of the traits and positively correlated with others.

Histogram:

Let's examine the Graph. It also supports the requirement for scalability by displaying the distribution of each characteristic and label over several ranges. Subsequently, discrete bars indicate that each of these variables is a categorical variable. Prior to using machine learning, we must address these categorical factors. There are two groups for our outcome labels: 0 for no disease and 1 for disease.

Bar Plot of Outcome:



The data is skewed toward datapoints with an outcome value of 0, which indicates that diabetes was not present in reality, as the preceding graph illustrates.

There are nearly twice as many people without diabetes as there are people with the disease.

k-Nearest Neighbors

Perhaps the most straightforward machine learning method is the k-NN algorithm. All that is required to build the model is to save the training data set. The method locates the "nearest neighbors," or closest data points, in the training data set in order to forecast a new data point.
Let's first examine if the relationship between model complexity and accuracy can be verified.
The graphic above displays the accuracy of the training and test sets on the y-axis in relation to the n_neighbors parameter on the x-axis. The prediction on the training set is flawless if we select just one nearest neighbor. However, the training accuracy decreases when additional neighbors are taken into account, suggesting that relying solely on the single nearest neighbor results in an overly complicated model. Around nine neighbors is where the finest performance occurs.

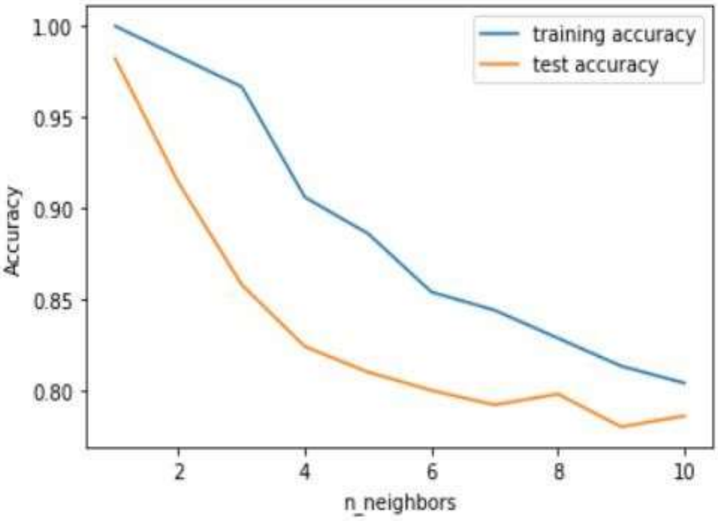


Table1: Accuracy in KNN

Training Accuracy	0.81
Testing Accuracy	0.78

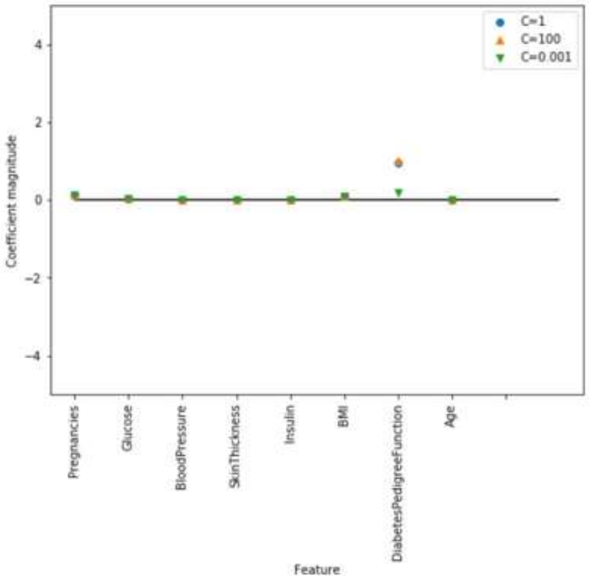
Logistic regression:

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

Table2: Accuracy in Logistic Regression

- ➢ Table 2 shows that 77% accuracy on the training set and 78% accuracy on the test set are obtained with the default value of C=1 in the first row, while 78% accuracy on both the training and test sets are obtained with C=0.01 in the second row.
- ➢ A more complicated model and less regularization may not generalize better than the default choice, as demonstrated by the slightly lower accuracy on the training set and slightly higher accuracy on the test set when C=100 is used.

✓ Therefore, we should choose default value C=1.



Decision Tree:

This classifier assigns the class values to each data point by building a decision tree.
Here, we may choose the maximum amount of characteristics that will be taken into account while building the model.

Training Accuracy	1.00
Testing Accuracy	0.99

Table3: Accuracy in Decision Tree

✓ Accuracy on the training set is 100% and the test set accuracy is also good.

The Value of Features in Decision Trees

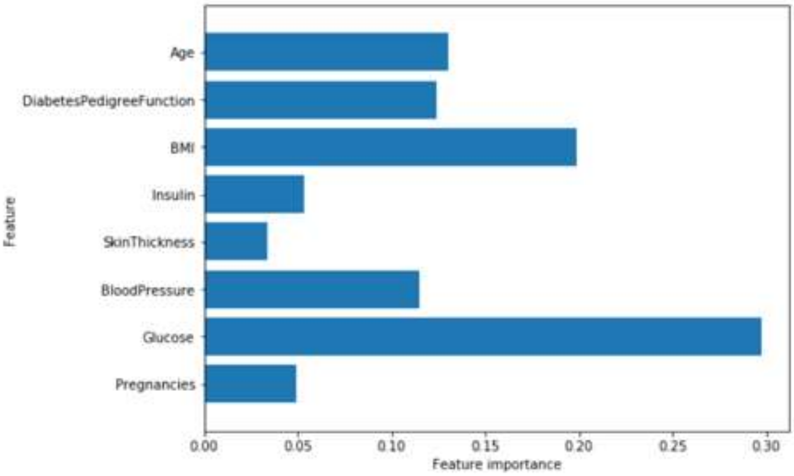
-The significance of each feature for a tree's decision is rated by feature importance.

-Each attribute is assigned a value between 0 and 1, where 0 denotes "never used" and 1 denotes "perfectly predicts the target."

Feature "Glucose" is by far the most important feature.

Random Forest:

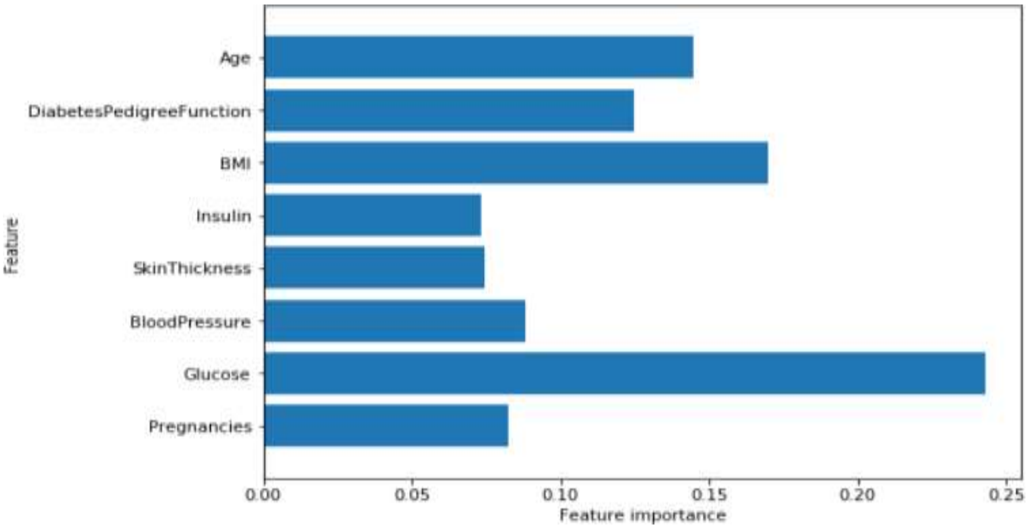
The idea of decision trees is advanced by this classifier. It produces a forest of trees, with each tree being made out of a random feature pick from the whole feature set.



Training Accuracy	**1.00**
Testing Accuracy	**0.974**

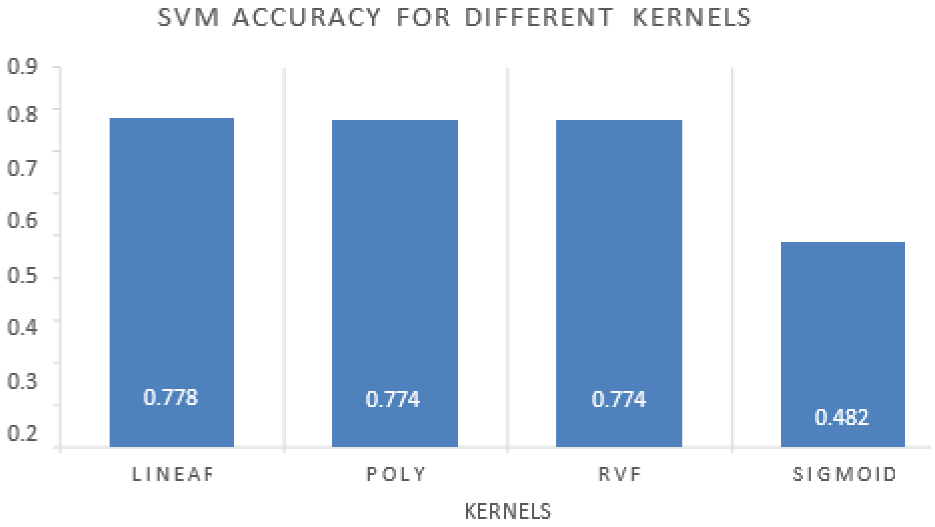
Table 4: Accuracy in Random Forest

The Value of Features in Random Forest



Support Vector Machine:

By varying the distance between the data points and the hyper plane, this classifier seeks to create a hyper plane that can effectively divide the classes. The hyper plane is chosen depending on a number of kernels. I experimented with four different kernels: sigmoid, linear, poly, and rbf. linear kernel performed the best for this dataset and achieved a score of 77%



Accuracy Comparison:

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

Table 5: Accuracy Comparison

This table shows Decision Tree algorithm gives the best accuracy with 98% training accuracy and 99% testing accuracy.

4.Conclusion

Early identification of diabetes is one of the major real-world medical challenges. This study aims to forecast diabetes by means of a system that is designed with methodical efforts. Five machine learning classification methods are examined and assessed using a variety of metrics in this paper. The John Diabetes Database is used to conduct experiments.

Using the Decision Tree method, experimental findings show that the developed system is adequate with an accuracy of 99%.

5.Future Work

The work may be expanded and enhanced for the automation of diabetes analysis, integrating some other machine learning algorithms, and the system that has been built with the machine learning classification algorithms can be used to predict or detect additional diseases in the future.

6. REFERENCES

[1]*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.*

[2]. 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.

[3]. 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.

[4]. 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.

[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]. Sharief, A.A., Sheta, A., 2014. Developing a Mathematical Model to Detect Diabetes Using Multigene Genetic Programming. International Journal of Advanced Research in Artificial Intelligence (IJARAI) 3, 54–59. doi:doi:10.14569/IJARAI.2014.031007.

[7]. Sisodia, D., Shrivastava, S.K., Jain, R.C., 2010. ISVM for face recognition. Proceedings - 2010 International Conference on Computational Intelligence and Communication Networks, CICN 2010 ,554– 559doi:10.1109/CICN.2010.109.

[8]. 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.

[9] Rebeiro, P.F.; Jenkins, C.A.; Bian, A.; Lake, J.E.; Bourgi, K.; Moore, R.D.; Koethe, J.R. Risk of incident diabetes mellitus, weight gain, and their relationships with integrase inhibitor-based initial antiretroviral therapy among persons with human immunodeficiency virus in the United States and Canada. Clin. Infect. Dis. ***2021****, 73, e2234–e2242. [****Google Scholar****] [****CrossRef****]*

[10] Berbudi, A.; Rahmadika, N.; Tjahjadi, A.I.; Ruslami, R. Performance of point-of-care testing compared with the standard laboratory diagnostic test in the measurement of HbA1c in Indonesian diabetic and nondiabetic subjects. J. Diabetes Res. ***2020****, 2020, 2037565. [****Google Scholar****] [****CrossRef****] [****PubMed****]*

[11] Jain, P.; Joshi, A.M.; Mohanty, S.P.; Cenkeramaddi, L.R. Non-invasive glucose measurement technologies: Recent advancements and future challenges. IEEE Access ***2024****, 12, 61907–61936. [****Google Scholar****] [****CrossRef****]*

[12]Lekha, S.; Suchetha, M. Recent advancements and future prospects on e-nose sensors technology and machine learning approaches for non-invasive diabetes diagnosis: A review. IEEE Rev. Biomed. Eng. ***2020****, 14, 127–138. [****Google Scholar****] [****CrossRef****] [****PubMed****]*

[13] Paleczek, A.; Rydosz, A. Review of the algorithms used in exhaled breath analysis for the detection of diabetes. J. Breath Res. ***2022****, 16, 026003. [****Google Scholar****] [****CrossRef****]*

[14] Monte-Moreno, E. Non-invasive estimate of blood glucose and blood pressure from a photoplethysmograph by means of machine learning techniques. Artif. Intell. Med. ***2011****, 53, 127–138. [****Google Scholar****] [****CrossRef****]*

[15] Paleczek, A.; Grochala, D.; Rydosz, A. Artificial breath classification using XGBoost algorithm for diabetes detection. Sensors ***2021****, 21, 4187. [****Google Scholar****] [****CrossRef****]*

[16]Wang, Z.; Wang, C. Is breath acetone a biomarker of diabetes? A historical review on breath acetone measurements. J. Breath Res. ***2013****, 7, 037109. [****Google Scholar****] [****CrossRef****]*

[17]Saasa, V.; Beukes, M.; Lemmer, Y.; Mwakikunga, B. Blood ketone bodies and breath acetone analysis and their correlations in type 2 diabetes mellitus. Diagnostics ***2019****, 9, 224. [****Google Scholar****] [****CrossRef****] [****PubMed****]*

[18]Güntner, A.T.; Weber, I.C.; Schon, S.; Pratsinis, S.E.; Gerber, P.A. Monitoring rapid metabolic changes in health and type-1 diabetes with breath acetone sensors. Sens. Actuators B Chem. ***2022****, 367, 132182. [****Google Scholar****] [****CrossRef****]*

[19]Yan, K.; Zhang, D.; Wu, D.; Wei, H.; Lu, G. Design of a breath analysis system for diabetes screening and blood glucose level prediction. IEEE Trans. Biomed. Eng. ***2014****, 61, 2787–2795. [****Google Scholar****] [****CrossRef****]*

*[20]Lekha, S.; Suchetha, M. Non-invasive diabetes detection and classification using breath analysis. In Proceedings of the 2015 International Conference on Communications and Signal Processing (ICCSP), Melmaruvathur, India, 2–4 April 2015; pp. 955–958. [****Google Scholar****]*