DIABETES PREDICTION USING MACHINELEARNING

Premala, Bhande[1*], Mohammad Kashif Ahteasham[2], Biradar Vaibhav[3], Srikanth. T[4]
Dept. of Computer Science and Engineering
Guru Nanak Dev Engineering College, Bidar, Karnataka, India
Visvesvaraya Technological University Belagavi Karnataka, India

Abstract - Diabetes is a chronic condition that could lead to a global health crisis. The International Diabetes Federation estimates that 382 million people worldwide have diabetes. By 2035, this will have doubled to 592 million. A condition known as diabetes is brought on by high blood glucose levels. The symptoms of increased thirst, hunger, and frequency of urination are brought on by this elevated blood glucose. One of the primary causes of renal failure, blindness, amputations, heart failure, and stroke is diabetes. Our bodies convert food into sugars, or glucose, when we consume. Our pancreas is meant to release insulin at that point. Insulin functions as a key to unlock our cells, letting glucose in and enabling us to utilise it as fuel. But this system is ineffective in the case of diabetes. Although type 1 and type 2 diabetes are the most common forms, there are other kinds as well, such as gestational diabetes, which appears during pregnancy. In data science, machine learning is a young scientific discipline that studies how machines pick up knowledge via experience. The purpose of this research is to develop a system that can more correctly detect diabetes in individuals at an early stage by combining the results of many investigations. Machine learning techniques. There is usage of methods such as K closest neighbour, decision tree, random forest, logistic regression, and support vector machine. It is computed what the model's accuracy is while employing each of the algorithms. The model that predicts diabetes isthen chosen based on the best accuracy then chosen based on the best accuracy. The algorithms. The model that predicts diabetes is then chosen based on the best accuracy. The model's accuracy is while employing each of the algorithms. The model that predicts diabetes is then chosen based on the best accuracy.is usage of methods such as K closest neighbour, decision tree, random forest, logistic regression, and support vector machine. It is computed what the model's accuracy is while employing each of the algorithms. The model that predicts diabetes is then chosen based on the best accuracy.pregnancy.In data science, machine learning is a young scientific discipline that studies how machines pick up knowledge via experience.The purpose of this research is to develop a system that can more correctly detect diabetes in individuals at an early stage by combining the results of many investigations.Machine learningtechniques. There is usage of methods such as K closest neighbour, decision tree, random forest, logistic regression, and support vector machine. It is computed what the model's accuracy is while employing each of the algorithms. The model that predicts diabetes is then chosen based on the best accuracy.

Keywords: Decision tree, K closest neighbour, machine learning, diabetes, logistic regression, support vector machines, and accuracy.

1 INTRODUCTION

Even in young individuals, 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. The source of sugar (glucose) is the dietary items, particularly those 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 substances that we ingest into glucose. 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 transferred to our cells.

Types of Diabetes

Type 1 diabetes means that the immune system is compromised and the cells fail to produce insulin insufficient amounts. There are no eloquent studies that prove the causes

of type 1 diabetes and there are currently no known methods of prevention.
Type 2 diabetes means that the cells produce a low quantity of insulin or the body can't use the insulin correctly. This is the most common type of diabetes, thus affecting 90% of persons diagnosed with diabetes. It is caused by both genetic factors and the manner of living.

Gestational diabetes appears in pregnant women who suddenly develop high blood sugar. In two thirds of the cases, it will reappear during subsequent pregnancies. There is a great chance that type 1 or type 2 diabetes will occur after a pregnancy affected by gestational diabetes.

Symptoms of Diabetes

- Frequent Urination
- Increased thirst
- Tired/Sleepiness
- Weight loss
- Blurred vision
- Mood swings
- Confusion and difficulty concentrating
- frequent infections

Causes of Diabetes

Genetic factors are the main cause of diabetes. It is caused by at least two mutant genes in the chromosome 6, the chromo some that affects there sponse of the body to various antigens. Viral infection may also influence the occurrence of type 1 and type 2 diabetes. Studies have shown that infection with viruses such as rubella, Coxsackie virus, mumps, hepatitis Bvirus, and cytomegalovirus increase the risk of developing diabetes.

2 LITERATUREREVIEW

To determine if a person has diabetes or not, Yasodhaet al.[1] 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 nine different properties. The two categories that are being discussed in these occurrences of the dataset are blood tests and urine testing. The implementation in this study may be carried out by classifying the data using WEKA, and the data is evaluated using the 10-fold cross validation technique, which works very well on tiny datasets, and the results are compared. Among them are the J48, Random Tree, REP Tree, and Naïve Bayes. The results showed that, among other things, J48 performs the best, with an accuracy of 60.2%.Aiswarya et al. [2] use Decision Tree and Naïve Bayes algorithms in classification analysis to look for patterns in the data that lead to the detection of diabetes in order to find solutions. The goal of the study is to provide a quicker and more effective way to diagnose illnesses, which will enable patients to receive treatment on schedule. The study found that the naïve Bayes algorithm yields an accuracy of 79.5% when utilising a 70:30 split, whereas the J48 algorithm yields an accuracy rate of 74.8% when using the PIMA dataset and cross validation technique. In order to compare and analyze the outcomes of multiple classification methods in WEKA, Gupta et al. [3] attempted to find and calculate the accuracy, sensitivity, and specificity percentage of various methods. Additionally, the study compared the performance of the same classifiers when implemented on some other tools, such as Rapidminer and Matlab, using the same parameters (i.e. accuracy, sensitivity, and specificity). 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 Rapid minner. The diabetes dataset is the main subject of Lee et al.'s [4] use of the CART decision tree algorithm.

The diabetes dataset is the main subject of Lee et al.'s [4] use of the CART decision tree algorithm. Following the data's application of the resample filter. The author places a strong emphasis on the issue of class imbalance and how it must be resolved before using any algorithms to increase accuracy rates. Class imbalances are most commonly found in datasets with dichotomous values, which indicate that the class variable has two alternative

outcomes. If this imbalance is identified early in the data preprocessing step, it may be easily handled and will improve the prediction model's accuracy.

3 METHODOLOGY

The several classifiers used in machine learning to predict diabetes will be covered in this section. We will also go over our suggested methods in an effort to increase accuracy. The present paper employed five distinct methodologies. 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 original source of the diabetes data set was https://www.kaggle.com/johndasilva/diabetes.

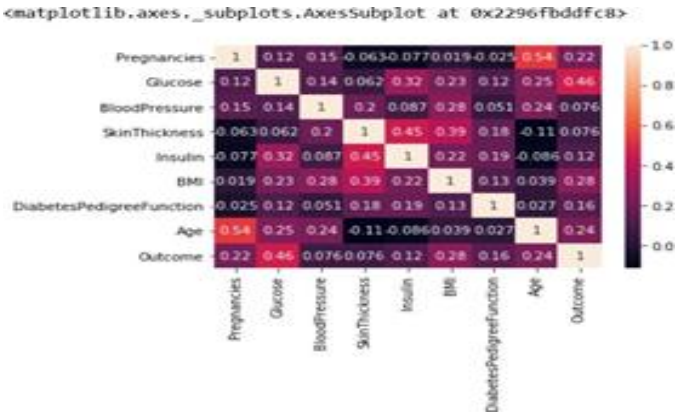
Diabetes dataset of two thousand instances. The goal is to determine whether or not the patient has diabetes by using the measurements.

	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

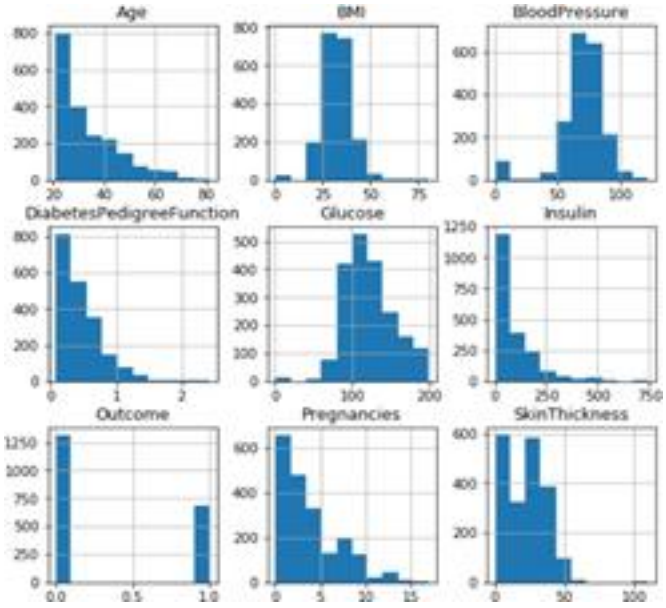
The diabetes data set consists of 2000 data points, with 9 features each.
“Outcome” is the feature we are going to predict, 0 means No diabetes, 1means diabetes.

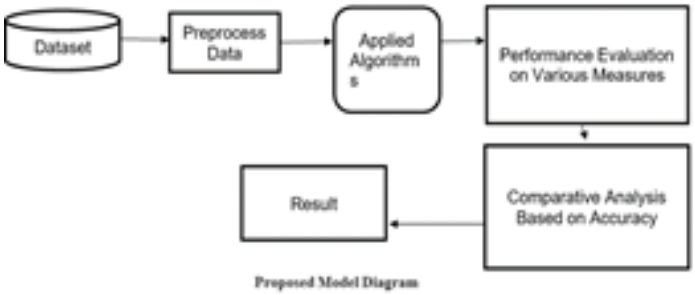
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 9 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Pregnancies              2000 non-null   int64
 1   Glucose                  2000 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
```

Correlation Matrix:



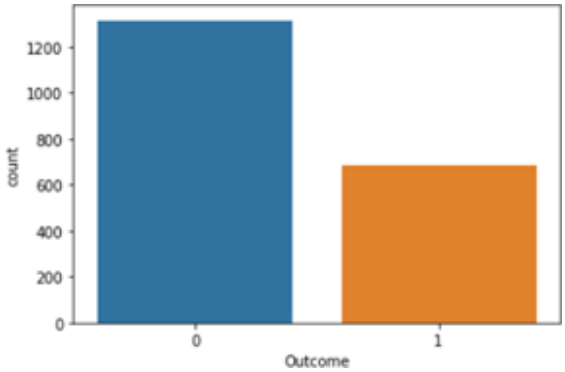
There is no null values in dataset.





Proposed Model Diagram

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.



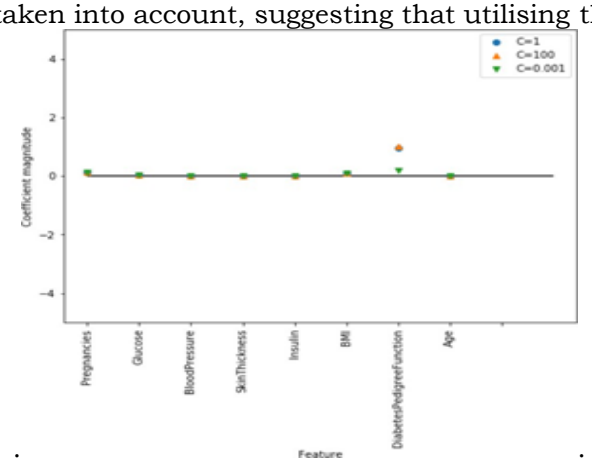
The graph above demonstrates how the data is skewed in favour of datapoints with an outcome value of 0, which indicates that diabetes was not genuinely present. There are about twice as many non-diabetics as there are people with diabetes.

K-Nearest Adjacents:

Perhaps the most straightforward machine learning algorithm is the k-NN algorithm. Storing the training data set is the only step involved in building the model. The method locates the closest data points—its "nearest neighbors"—in the training data set in order to predict a new data point.

Let's first see whether we can validate the relationship between model complexity and accuracy:

The plot above contrasts the setting of neighbors on the x-axis with the accuracy of the training and test sets on the y-axis. The prediction on the training set is flawless if we select the single nearest neighbour. However, the training accuracy decreases with additional neighbours taken into account, suggesting that utilising the single nearest



Neighbor results in an overly complex model. It is estimated that nine neighbours have the best performance.

Training Accuracy	0.81
Testing Accuracy	0.78

Table-1

Logistic regression:

Logistic Regression is one of the most common classification algorithms.

	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

First row: 77% accuracy on the training set and 78% accuracy on the test set are provided by the default setting of C=1.

Using C=0.01, the results in the second row show 78% accuracy on both the test and training sets.

Using C=100 yields somewhat lower accuracy on the training set and slightly higher accuracy on the test set, indicating that a more complex model with less regularization may not generalize more effectively than the default configuration.

As a result, C=1 should be our default value.

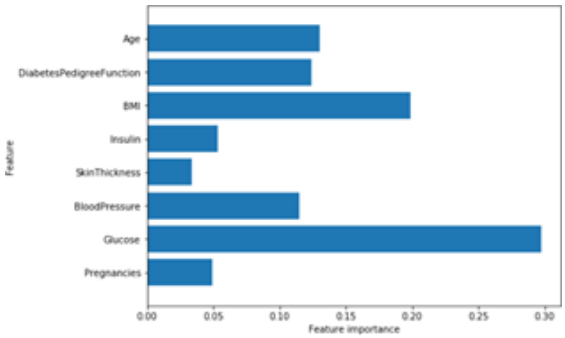
Decision Tree:

This classifier creates a decision tree based on which, it assigns the class values to each data point. Here, we can vary the maximum number of features to be considered while creating the model.

Training accuracy	1.00
Testing accuracy	0.99

Table 3

The accuracy on the training set is 100% and the test set accuracy is also good. Feature Importance in Decision Trees Feature importance rates how import an teach feature is for the decision a tree makes. It is a number between 0 and 1 for each feature, where 0 means “not use data all” and 1 means “perfectly predicts the target”.

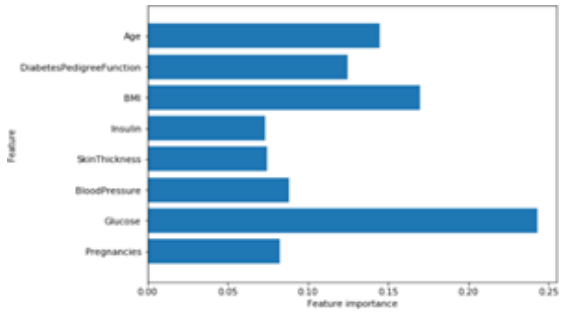


Feature “Glucose” is by far the most important feature.

Random Forest:

This classifier takes the concept of decision trees to the next level. It creates a forest of trees where each tree is formed by a random selection of features from the total features.

Training accuracy	1.00
Testing accuracy	0.974

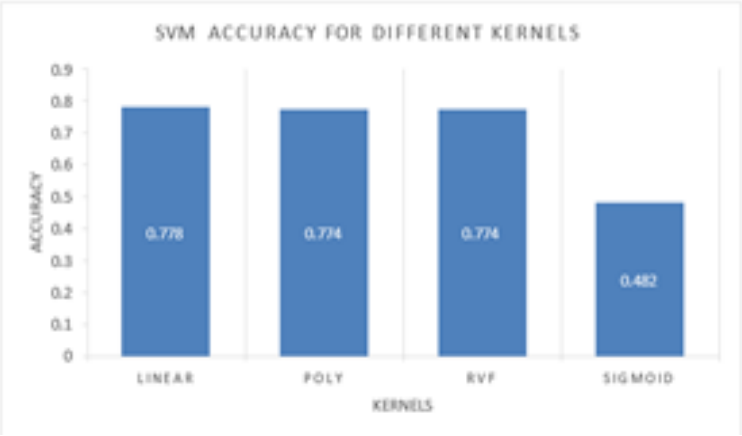


Feature importance in Random Forest

Like the single decision tree, the random forest similarly ranks the "Glucose" feature highly, but it also ranks the "BMI" feature as the second-most informative feature overall.

Helping Vector Computer:

By changing 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 using the linear, poly, rbf, and sigmoid kernels.



As can be seen from the plot above, the linear kernel performed the best for this data set and achieved as core 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

Table 5 Shows the accuracy values for all five machine learning algorithms.
Table - 5 shows that Decision Tree algorithm gives the best accuracy with 98% training accuracy and 99% testing accuracy.

V. CONCLUSION AND FUTUREWORK

Early identification of diabetes is one of the major real-world medical issues. In this work, a methodical approach to developing a system that predicts diabetes is taken. In the process, five machine learning classification methods are examined and assessed using different metrics. Research is conducted using the **pima Diabetes Database**. The effectiveness of the planned system is assessed by experimental findings using an 99% accuracy was attained with the Decision Tree algorithm.

In the future, additional diseases may be predicted or diagnosed using the system that was created and the machine learning classification algorithms that were employed. The work can be expanded upon and enhanced to include additional machine learning techniques and automate the analysis of diabetes.

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 ofKingSaudUniversity-ComputerandInformationSciences25,127–136.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 ComputerApplications54,21–25.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 Pimaindi and diabetes dataset using naive bayes with genetic algorithms 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 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, 554–559 doi:10.1109/CICN.2010.109.
8. Sisodia, D., Singh, L., Sisodia, S., 2014. Fast andAccurate Face Recognition Using SVM and DCT, in: Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS2012),

December 28-30, 2012, Springer. pp. 1027–1038.
9. https://www.kaggle.com/johndasilva/diabetes
10. Rani, A. S., & Jyothi, S. (2016, March). Performance analysis of classification algorithms under different data sets. In Computing for Sustainable Global Development (INDIA Com), 2016 3rd International Conference on (pp. 1584-1589)