JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Diabetes Prediction Using Supervised Machine Learning Algorithms

1*Dr.E.Srinvas Reddy, 1 Dean
2 Minda Sandhya, 2 BTech Student
3 Chintada Akhil, 3 BTech Student
4 Maddineni Sirisha, 4 BTech Student
5 Palaka Sagar 5 BTech Student

1* Department of Computer Science and Engineering
1* Acharya Nagarjuna University, ANU College of Engineering and Technology, Guntur, Andhra Pradesh.

ABSTRACT:

Diabetes is a common health problem that is typified by

consistently elevated blood sugar levels, especially in Bangladesh. Heart attacks, strokes, kidney failure, and blindness are just a few of the major health issues it causes. The ability to take potentially life-saving action and intervene promptly is made possible by early detection. Regretfully, diabetes is becoming more and more common. The usage of this work is to analyze the predictive many

widely used machine learning algorithms for diabetes. Technological developments in machine learning have yielded substantial benefits for the medical

industry by providing a broad spectrum of algorithmic approaches. In this study,

six popular machine learning approaches are employed to analyze performance measures.

Keywords: supervised machine learning algorithms are SVM, Logistic Regression, Random Forest, Decision tree, KNN, and Navie Bayes theorem.

1.INTRODUCTION:

Finding people who are susceptible to this chronic illness before it manifests itself is the first step in predicting diabetes. For diabetes and its related problems to be prevented or delayed in progression, early detection is essential. Prediction techniques make use of a variety of variables, including lifestyle, health measures, and genetics. Healthcare practitioners can evaluate these characteristics to identify high-risk individuals by utilizing sophisticated technologies like statistical models

and supervised machine learning algorithms . The main usage of this project is to improve the patient outcomes and we predicted the diabetes early with the help of their causes.

They are three types of diabtes.They are:

Type 1 diabetes: In type1 diabetes , Damage the immune system and attacks the islet cells in the pancreas that make the insulin. Because pancreas does not make insulin

In type2 diabetes: pancreas doesnot produce the insulin to all body cells so in that circumstances occur Type 2 diabetes.

Gestational diabetes: gestational diabetes is not caused by a lack of of insulin, but by other hormones produced during pregnancy that can make insulin less effective, a condition referred to as insulin resistance.

2.LITERATURE REVIEW:

A brand-new approach to diagnosing and categorizing diabetes has been presented by researchers. Patients' blood sugar levels are monitored on a regular basis with continuous glucose monitoring (CGM). Chinese population data was

examined through data analysis of clinic records at the People's Hospital of China. They created a new indicator for diabetes diagnosis and classification by using an AdaBoost variant algorithm to extract 17 attributes, and they tested it with 90.3% accuracy. The average conversation duration (ACC) and mean conversation content (MCC) were among the metrics utilized to analyze the data. and In [2] they found the accuracy 98.35% compared to other accuracies. and another researcher said that AdaBoost Classifier produced 98.8%.

3. PROPOSED METHODOLOGY:

Because machine learning algorithms are limited to processing numerical input, converting nominal data to numerical data poses a substantial barrier. These algorithms can neither directly measure nor make use of text data in its original format. For this reason, in order to guarantee accuracy in our study, it is essential to convert textual information into a quantitative format.

We first processed and transformed the nominal data into numerical form in order to address this. Our algorithms then made advantage of these pre-processed data. To accomplish this conversion and the analysis that followed, we used supervised machine learning techniques. we check the null values with the help of isnull method and removed the null values using mean, mode, median after insert the dataset. we found the accuracy 99% of Random forest and next highest accuracy 98% of SVM.

we found SVM and Random Forest algorithms are best algorithms.

A. Data Collection & Pre-Processing

When working with missing, noisy, or inconsistent data, data preparation is the crucial step in the data mining . To make sure the data is presented in a consistent and appropriate way, this step includes a number of activities, such asdata conversion, processing, integration, defuzzification, and cleansing.

We used a diabetes dataset from the UCI repository, which has 17 variables linked to patient and hospital outcomes, for our case study. The purpose of this dataset is to assess how well ensemble algorithms anticipate future events. It consists of standard treatment information.

B. Data Description

We began by analyzing the structure of our dataset, comprising 17 columns and a total of 521 records. To identify missing values, we utilized data visualization techniques. Subsequently, we visualized each column separately to determine the

distribution of True and False classes, presenting both the percentage and absolute counts of data for each column.

1. **Age:** The people's ages in the dataset.
2. **Gender:** Generally classified as either male or female, the individuals' gender.
3. **Polyuria:** This indicates whether or not the people urinate excessively, or polyuria.
4. **Polydipsia:** Indicates whether the people have excessive thirst, or polydipsia.
5. **Sudden Weight Loss:** This indicates whether or not the people have lost weight suddenly.
6. **Weakness:** This shows whether a person exhibits weakness.
7. **Polyphagia:** Indicates whether the people have polyphagia, or an insatiable appetite.
8. **Genital Thrush:** This denotes the presence of a fungal infection called genital thrush in the individuals.
9. **Visual Blurring:** This indicates whether or not people report visual blurring.
10. **Itching:** Provides information about whether or not the people are itchy.
11. **Irritability:** This indicates if the people are easily agitated.
12. **Delays in Healing:** Shows Whether the people

C. Proposed Model Workflow

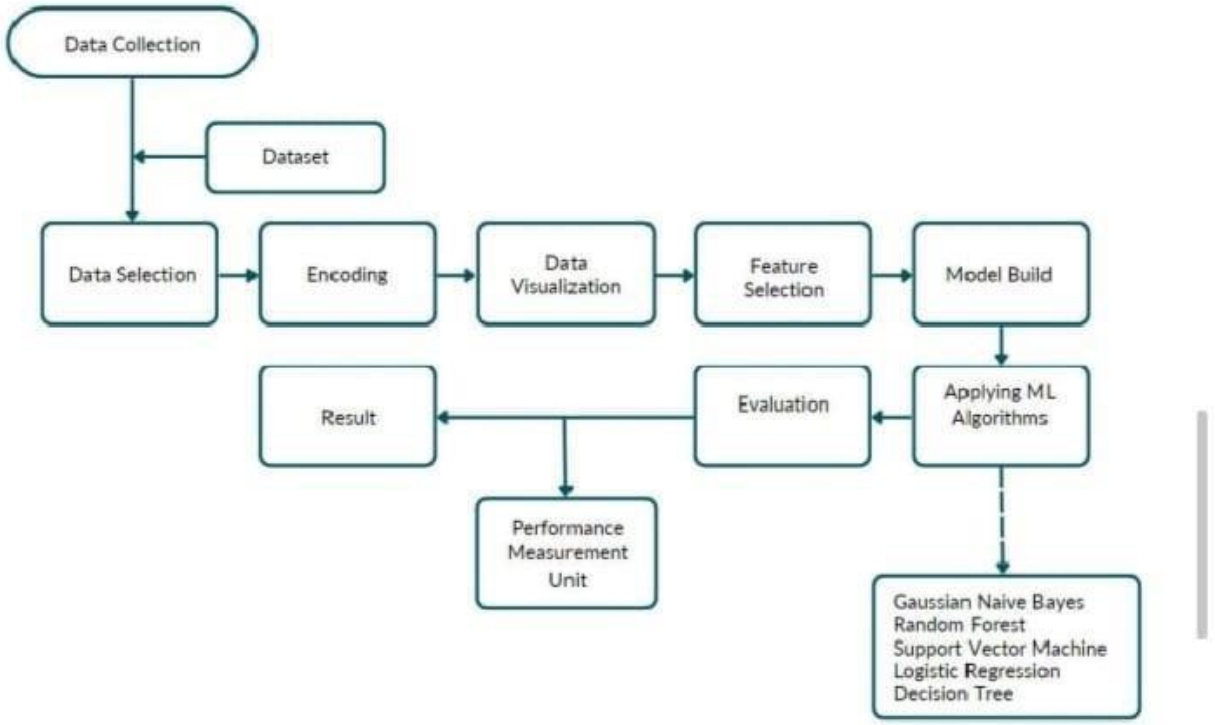


Fig. 1. Proposed model workflow

D. Machine Learning Model

A machine learning model is a one type of program that can find outs the patterns or make decisions from a previously unseen dataset. A machine learning model is used for whether the person is diabetic or not . A machine learning model is a graphical or mathematical expression of an algorithm. machine learning model is applied on machine learning algorithms.

1) Random Forest (RF)

In machine learning, Random Forest is a supervised machine learning method this is well-known for working well in both regression and classification problems. To improve model performance and solve challenging problems, it make use of group learning, which uses several classifiers. The way Random Forest works is that it builds a lot of decision trees, and the density of the forest has a big impact on the outcome. In general, more trees in the forest produce more trustworthy outcomes. Bagging, which is a prominent characteristic of decision tree algorithms,is included into Random Forest. Even though Random Forest is self- governing, it usually needs human oversight to guarantee peak performance.

2) Logistic Regression (LR)

Based on a given dataset, a statistical approach called logistic regression (LR) is used to predict binary outcomes, such as yes or no. It finds at the connection between a dependent variable and independent variable. A political
candidate's success or failure in an election or a high school senior's admittance to a university are two examples of situations where logistic regression might be used. By providing distinct binary options, this method streamlines decisionmaking. Under supervised learning, more precisely in the context of classification,
lies logistic regression. The link between the input variable x and the feature variable y in classification issues is clearly discrete.

3) Decision Tree (DT)

For problems involving prediction and classification, Decision Trees (DT) are an extremely useful tool. every node in a decision tree presents a test on an attribute; the branches in a decision tree indicate the test results; and the leaf nodes, also known as terminal nodes, represent the class labels. For data categorization and predictive modeling, decision trees are very helpful because of this. Every hypothesis is presented by a node in the tree, and the expected value is indicated by the endpoints of the branches. Decision trees can be organized in a variety of ways, depending on the information and issue at hand. When there are lots of data for model training and few classes, they function well. They might have trouble, though, if there are a lot of classes and not a lot of training data.

3) Support Vector Machine

SVM's flexible performance and adaptability make it useful in a long distance of application, include finance, bioinformatics, images and text categorization, and bioinformatics. However, the effectiveness of SVM is largely dependent on choosing an appropriate kernel function and optimizing its parameters, both of which can be difficult. Despite this, SVM is a useful tool in the machine learning toolbox since, when used properly, it usually yields better accuracy and generalization performance. we used three kernals in SVM algorithm they are linear, rbf, poly, sigmoid.

4) K-Nearest Neighbor

KNN is effective non-parametric approach used in regression and classification tasks. It operates under the principle that the class of a particular data is determined by which is the majority class among its K closest

neighbor in the feature space.
The algorithm follows these steps: for each instance, calculate the distance from every data point to another data point then we find out some distance after getting the distance we can decide whether the data point goes into class 1 or class 2 .
Distance Metric: While various distance metrics can be utilized, Euclidean distance is commonly preferred, although Manhattan distance is another option.
Hyperparameter: The number of neighbors (K) is a crucial hyperparameter that significantly impacts the algorithm's performance. Proper selection of K is essential for achieving optimal results.

5) Gaussian NB

An technique called Gaussian Naive Bayes (Gaussian NB) is based on the statistical Bayes theorem and is used to create probabilistic classifications by examining the correlations between features in a dataset. In order to use this method, one must first assume that a class variable's value is independent of every other property in the dataset. Under the premise that all attribute's characteristics have an equal and uncorrelated impact on the variable's output value, Gaussian NB functions. The data are categorized into appropriate classes prior to calculating the variance and mean of continuous attribute values (x).

4. . EXPERIMENTAL RESULT

Mistakes are frequent while teaching or extrapolating results. It is well known that when model complexity increases, training error rates tend to go down, which can aid in lowering training errors. Better generalization isn't always the result of this, though. Reducing training mistakes may not always enhance the model's performance on unobserved data, as the Bias-Variance Decomposition (Bias + Variance) method illustrates. When training error decreases and test error rises as a result, overfitting takes place.

Several techniques, such as accuracy, precision, recall, f1-score can be used to assess a classification system. The performance of the models was measured by the authors using a variety of techniques. While some research included several

measures for assessment, others concentrated on just one. The model's efficacy was evaluated in this study.

first fall model is tested and after tested we train the model then it gets accuracy from all algorithms.

we improved the performance of model using accuracy, precision, recall, f1_score.

1.Accuracy

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

2.Precision

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

3.Recall

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$$

4.F1-Score

$$\text{F1-Score} = 2\times\frac{\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}$$

where:

Cases that are accurately predicted as positive (diabetes) are known as TPs (True Positives).

-True Negatives, or TNs, are the cases where the outcome is accurately expected to be negative (diabetic).

-False positives, or FPs, are situations where a positive result is accurately predicted (e.g., hypothesized to be diabetic when in fact it is not).

-False Negatives, or FNs, are cases that are accurately forecasted as negative but are actually diagnosed as diabetes.

Final Result:

Dataset was trained using six different machine learning techniques. These algorithms differed slightly from one another. These accuracy values show how well each algorithm performs in accurately predicting situations that are neither diabetes nor non-diabetic. With 99% accuracy, Random Forest was the most

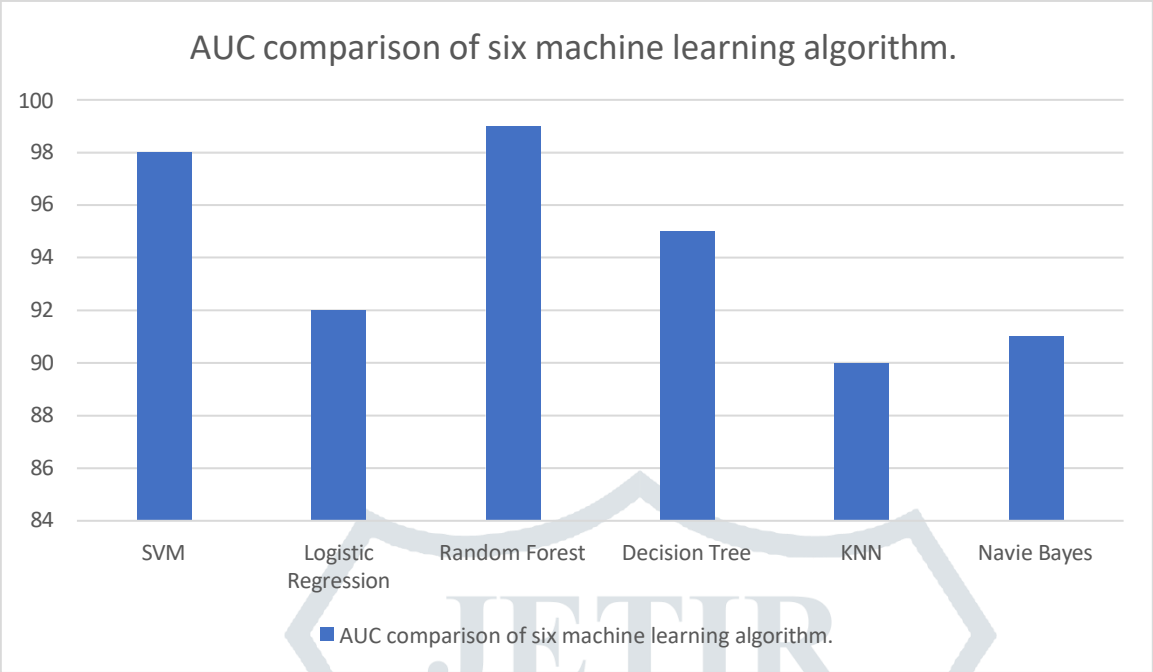
accurate, closely followed by SVM with 98%. accuracy of 90% KNN was the least accurate algorithm on the list. only which algorithm give more accuracy that algorithm is considered to find out the whether the person is diabetes or not.

A.Result

The dataset was trained using six different machine learning techniques. These algorithms differed slightly from one another. These accuracy values show how well each algorithm performs in accurately predicting situations that are neither diabetes nor non-diabetic. With 99% accuracy, RandomForest was the most accurate, closely followed by SVM with 98%. With an accuracy of 90%, KNN was the least accurate algorithm on the list.

TABLE. Accuracy of this model

Algorithm	**Test Accuracy**	**Train Accuracy**
SVM	94%	98%
Logistic Regression	93%	92%
Random Forest	100%	99%
Decision Tree	100%	95%
KNN	95%	90%
Navie Bayes	89%	91%



Confusion Matrix

A simple and quick way to summarize a classification system's efficacy is to utilize confusion matrices. The dataset only has two categories, but if there is a significant variation in the number of observations within or across categories, the classification may be incorrect. One can compute a confusion matrix (CM) to gain more insight into the classification method's accuracy.
confusion matric is find the errors in model and improve the performance.

Confusion Matrix of SVM

$$\begin{bmatrix} 28 & 5 \\ 4 & 67 \end{bmatrix}$$

Confusion Matrix of Logistic Regression

$$\begin{bmatrix} 28 & 5 \\ 4 & 67 \end{bmatrix}$$

Confusion Matrix of Random Forest

$$\begin{bmatrix} [28 & 5 \\ 4 & 67] \end{bmatrix}$$

Confusion Matrix of Decision Tree

$$\begin{bmatrix} [28 & 5 \\ 4 & 67] \end{bmatrix}$$

Confusion Matrix of KNN

$$\begin{bmatrix} [28 & 5 \\ 4 & 67] \end{bmatrix}$$

Confusion Matrix of Naïve Bayes

$$\begin{bmatrix} [28 & 5 \\ 4 & 67] \end{bmatrix}$$

5. DISCUSSION

Diabetes is one of the most frequent diseases that has a serious negative influence on health. Diabetes, which is characterized by increased blood sugar levels, is frequently mentioned as a contributing factor to a number of other health conditions. The bloodstream's glucose provides the body with its main energy source. The most dangerous side effect of diabetes is blindness. Transporting glucose from the bloodstream into cells for energy requires the hormone insulin, which is generated by the pancreas.

Diabetes and its after effects, such as blindness, currently impact a sizable section of the world's population. Several health consequences can form when the body's stability to produce or use insulin is compromised.

6. CONCLUSIONS AND FUTURE WORK

SVM and Random Forest algorithms are produce high accuracy compared to all algortithms because SVM gives high accuracy because it have hyperplane that hyperplane maximizes the margin between different two classes, SVM is effective in high dimensional spaces and it can handles non linear data through kernel functions and this algorithm is well suited for long data sets.

Random forest also produce high accuracy because It have many decision trees then long data set divided several decision trees and we find out the majority of leaf node only that output come is considered. Reduce the overfitting and handle complex data sets.

There is a great deal of promise for bettering healthcare outcomes in Bangladesh through the use of machine learning algorithms for early predict the diabetes.

The Random Forest and SVM are shows to be the most accurate method by the research findings, allowing medical practitioners to recognize individuals who are at-risk and initiate early therapies.

To guarantee the responsible use of these technologies, it is imperative to address ethical issues like bias and data privacy. Any approach targeted at reducing the diabetes epidemic should also include the promotion of healthy lifestyles and preventative measures.

In summary, this research shows how machine learning techniques can improve patient outcomes and further wider public health objectives. In the future, research should concentrate on:

* Improving the accuracy of information on the prognosis of breast cancer.
* constructing more sophisticated classification schemes.
* better deep learning models are put into practice.
* Increasing these methods' precision and effectiveness

7.References

[1] J. Abdollahi and B. Nouri-Moghaddam, “Hybrid stacked ensemble combined with genetic algorithms for diabetes prediction,” Iran Journal of Computer Science 2022 5:3, vol. 5, no. 3, pp. 205–220, Mar. 2022, doi: 10.1007/S42044-022-00100-1.

[2] S. Islam Ayon and M. Milon Islam, “Information Engineering and Electronic Business,” Information Engineering and Electronic Business, vol. 2, pp. 21–27, 2019, doi: 10.5815/ijieeb.2019.02.03.

[3] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," Procedia Comput Sci, vol. 165, pp. 292– 299, Jan. 2019, doi: 10.1016/J.PROCS.2020.01.047.

[4] M. Saberi-Karimian et al., "Data mining approaches for type 2 diabetes mellitus prediction using anthropometric measurements," J Clin Lab Anal, vol. 37, no. 1, p. e24798, Jan. 2023, doi: 10.1002/JCLA.24798.

[5] M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryanigrum, and R. Yunanda, "Diabetes prediction using supervised machine learning," Procedia Comput Sci, vol. 216, pp. 21–30, Jan. 2023, doi: 10.1016/J.PROCS.2022.12.107.

[6] G. Cappon, M. Vettoretti, G. Sparacino, and A. Facchinetti, "Continuous Glucose Monitoring Sensors for Diabetes Management: A Review of Technologies and Applications," Diabetes Metab J, vol. 43, no. 4, pp. 383–397, Aug. 2019, doi: 10.4093/DMJ.2019.0121.

[7] J. Yang et al., "Modifiable risk factors and long term risk of type 2 diabetes among individuals with a history of gestational diabetes mellitus: prospective cohort study," BMJ, vol. 378, Sep. 2022, doi: 10.1136/BMJ-2022-070312.

[8] R. Srivastava and R. K. Dwivedi, "A Survey on Diabetes Mellitus Prediction Using Machine Learning Algorithms," Lecture Notes in Networks and Systems, vol. 321, pp. 473–480, 2022, doi: 10.1007/978-981-16-5987-4_48/COVER.

[9] M. R. Rajput and S. S. Khedgikar, "Diabetes prediction and analysis using medical attributes: A Machine learning approach", doi: 10.37896/JXAT14.01/314405.

[10] C. Y. Chou, D. Y. Hsu, and C. H. Chou, "Predicting the Onset of Diabetes with Machine Learning Methods," Journal of Personalized Medicine 2023, Vol. 13, Page 406, vol. 13, no. 3, p. 406, Feb. 2023, doi: 10.3390/JPM13030406.