International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Diabetes prediction by using machine learning algorithms

S.Umamaheswararao[1] *,B.Sriramya*[2]*, K.Prathyusha*[3]*, K.Siddhartha*[4]*, P.Mahesh*[5]

ECE Department
Aditya Institute Of Technology And Management,
Kotturu, Tekkali

ABSTRACT:

With 537 million cases worldwide, diabetes is the most common and fatal non-communicable disease. Diabetes can be brought on by a variety of causes, including an unhealthy diet, high blood pressure, abnormal cholesterol, a family history of the disease, and inactivity. One of this disease's most prevalent symptoms is increased urination. Long-term diabetics may experience a number of issues, including diabetic retinopathy, renal disease, nerve damage, and heart problems. However, if it is anticipated early on, the risk can be decreased. The use of machine learning classifiers to forecast the onset of diabetes based on different patient characteristics is investigated in this work. A dataset comprising lifestyle, clinical, and demographic information is used to train and assess many classification algorithms in order to determine which model performs best. The findings show how machine learning can effectively estimate the risk of diabetes, providing important information for early intervention and individualised treatment plans.

INTRODUCTION :

Among the world's nasty diseases is diabetes. Diabetes brought on by obesity, elevated blood sugar, and other factors. It affects the insulin hormone, which causes improper crab metabolism and raises blood sugar levels. When the body does not produce enough insulin, diabetes develops. About 422 million people worldwide, mostly in low- and middle-income nations, suffer from diabetes, according to the World Health Organisation (WHO). And by the year 2030, this might rise to 490 billion. Nonetheless, diabetes is prevalent in many nations, including China, India, and Canada. With over 100 million people living in India now, there are actually 40 million diabetics living there. One of the leading causes of death worldwide is diabetes.Early diagnosis and treatment of diseases like diabetes can save lives.

In order to achieve this, this paper investigates diabetes prediction using multiple characteristics associated with diabetes.

In order to forecast diabetes, we employ the Pima Indian Diabetes Dataset and a variety of machine learning classification and ensemble techniques. One technique for explicitly training computers or other machines is machine learning. Using gathered datasets, several machine learning techniques develop ensemble and classification models that efficiently gather knowledge.

Diabetes can be predicted with the help of such gathered data. Many machine learning techniques are capable of making predictions, but selecting the most effective method can be challenging. In order to make predictions, a variety of datasets are analysed using advanced algorithms that take into account important characteristics including age, BMI, blood pressure, glucose levels, and family history.

From conventional techniques like logistic regression to sophisticated deep learning neural networks, machine learning models are trained on historical data to identify complex patterns and associations suggestive of the risk of diabetes.

Making sure the models are interpreted correctly is one of the main problems with diabetes prediction, especially in the healthcare setting where openness is essential. Healthcare providers' trust is increased by interpretable models, which also offer insights into the variables influencing forecasts and facilitate well-informed decision-making.

In addition, it is critical to eliminate any potential biases in the data and models in order to guarantee just and equitable results, particularly when working with a variety of patient populations. The diabetes prediction process incorporates privacy protection, regulatory compliance, and ethical considerations as essential elements.

The ultimate purpose of machine learning-based diabetes prediction is to provide timely and accurate information to healthcare practitioners, allowing for tailored therapies and better patient outcomes. Modern algorithms help to change the way that healthcare is provided by utilising data to drive proactive approaches to public health and diabetes treatment. The potential for early diabetes detection at the nexus of health sciences and technology is enormous, and it will strengthen and adapt the healthcare system.

2.Overview machine learning:

2.1 Supervised Learning:

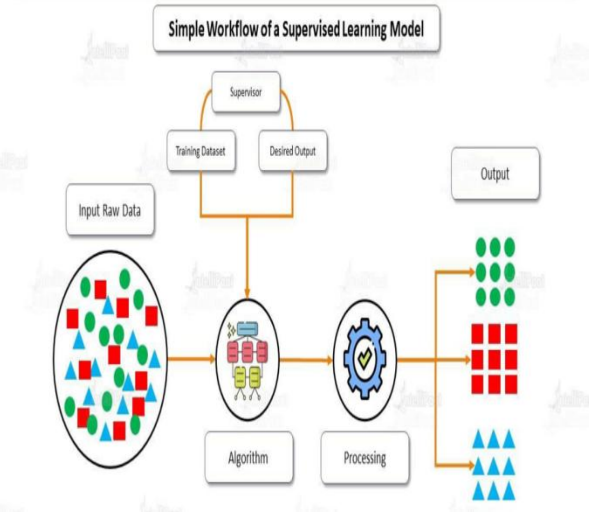
Supervised learning refers to the kind of machine learning wherein machines are trained using properly "labelled" training data, and then they make output predictions based on that data. Some input data has already been tagged with the appropriate output, as indicated by the labelled data. In supervised learning, the machines' training data serve as the supervisor, teaching them to accurately predict the output.

It uses the same idea that a student would learn under a teacher's guidance. The process of giving the machine learning model accurate input and output data is known as supervised learning. The objective of an algorithm for supervised learning is to identify a mapping function that associates the input variable (x) with the output variable (y).

In the actual world,

How supervised learning works :

Supervised learning required training labelled data. In order to do classification, we need to first label the data and then use it to train in model to classify them in groups. In supervised learning, we train your model on a labelled dataset that means we both raw input data as well as its results. We split our data into a training dataset and test dataset where the training dataset is used to train our model whereas the test dataset acts as new data for predicting results or to see the accuracy of our model.

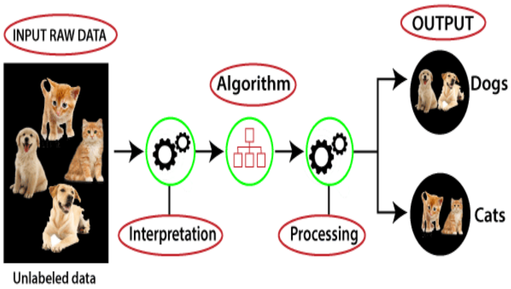


- The shape is designated as a square if it has four sides and each side is equal.
- The shape will be designated as a triangle if it has three sides.
- The shape will be called a hexagon if it has six equal sides. After training, we use the Est set to put our model to the test. The model's job is to recognize the shape. The computer has already been educated on a wide variety of shapes. When it encounters a new shape, it categorizes it based on several sides and forecasts the result.

Unsupervised Learning:

Unsupervised learning does not require labelled or classified data explicitly. In unsupervised learning, the information used to train is neither classified nor labelled in the dataset. Unsupervised learning studies on how systems can infer a function to describe a hidden structure from unlabelled data. The main task of unsupervised learning is to find patterns in the data.

How Unsupervised Learning Works :



In this case, the input data is unlabelled, meaning it is not categorized and no corresponding outputs are provided. The machine learning model is now fed this unlabelled input data in order to train it. It will first analyse the raw data to identify any hidden patterns before applying the appropriate techniques, including decision trees and k-means clustering, to the data.

After applying the appropriate algorithm, the algorithm groups the data objects based on the similarities and differences among them.

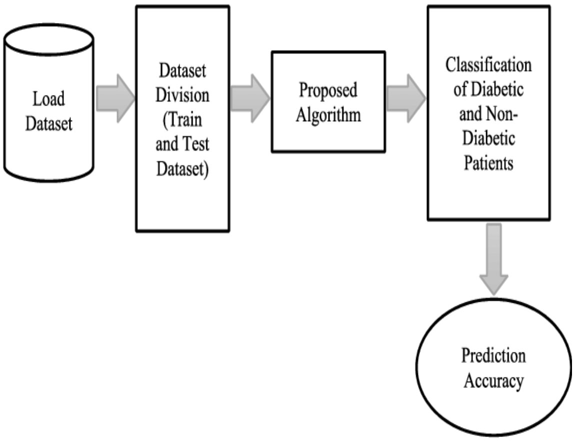
Problem statement:

Diabetes continues to be a major global health concern, affecting millions of people globally, despite advances in healthcare. In order to manage diabetes and avoid complications, early detection and prompt management are essential. The goal of this project is to create a reliable and accurate machine learning model that can be used to forecast a person's risk of developing diabetes based on their lifestyle, health-related characteristics, and demographics.

The intricacy of diabetes risk variables, the requirement for predictions that can be understood in a clinical context, and the possibility of dataset imbalance are the main issues that need to be resolved. In order to enable proactive and individualised preventative actions, the proposed model should offer trustworthy predictions that assist medical practitioners in identifying people who are at risk of getting diabetes.

Methodology:

1.Proposed Algorithm


Load Dataset
Dataset Division (Train and Test Dataset)
Proposed Algorithm
Classification of Diabetic and Non-Diabetic Patients
Prediction Accuracy

a) **Data Collection:** Compile a feature-rich, diversified dataset. Age, BMI, family history, blood pressure, cholesterol, and physical activity are typical characteristics.

b) **Data Preprocessing**:Remove outliers and missing values to clean up the dataset. To bring numerical properties to a same scale, normalise or standardise them. Use methods such as one-hot encoding to encode categorical variables.

c) **Feature Selection:** Use methods such as correlation analysis or feature importance from tree-based models to determine which features are significant. Select characteristics that significantly affect the prediction of diabetes.

d) **Model Selection**: Pick suitable machine learning algorithms. Neural networks, support vector machines, decision trees, random forests, logistic regression, and random forests are popular options for binary classification problems like diabetes prediction.

e) **Model Training:** Using the training dataset, train the chosen model. To enhance the model's performance, optimise the hyperparameters.

f) **Model Evaluation:** Use metrics like accuracy, precision, recall, F1 score, and area under the ROC curve (AUC ROC) to assess the model on the testing dataset. Cross-validate to make sure it's resilient.

g) **Interpretation of the Model:**
Recognise how the model generates forecasts. For example, in logistic regression, look at coefficients to see which features have the greatest impact on predictions.

h) **Validation and Fine Tuning:**
To make sure the model is generalizable, validate it using fresh, untested data. Based on the validation results, adjust the model as needed.

i) **Deployment:** After you're happy with the model's performance, use it in an actual environment. Put monitoring systems in place to guarantee accuracy over time.

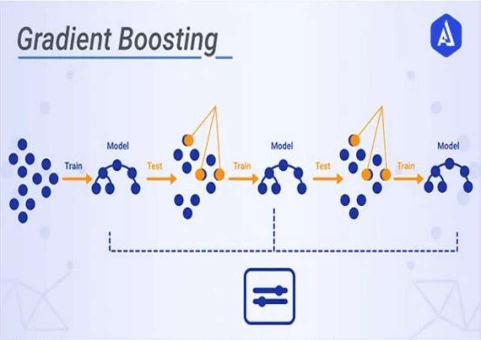
Classifiers

1.Logistic Regression
2.Random Forest
3.Decision Tree
4.KNN
5.SVM
6.Adaptive Boosting
7.Gradient Boosting

1. Logistic Regression:

A supervised learning approach called logistic regression is applied to binary classification tasks in which there are two possible outcomes for the target variable (class 0 or 1). To model the likelihood that a given input is a member of the positive class is the aim of logistic regression.

2. Random Forest:
   During training, a Random Forest algorithm creates a collection of decision trees, or forest, and then aggregates these forecasts to enhance accuracy and generalisation. After being trained on a random subset of the training set, each decision tree in the forest generates predictions on its own.
3. Decision Tree: A Decision Tree Classifier is a simple yet powerful machine learning algorithm used for classification tasks, including epileptic seizure recognition. It works by partitioning the feature space into smaller regions based on the values of input features, eventually making a decision about the class label for each region.
4. KNN: A non-parametric approach called K-Nearest Neighbours uses the majority class or average of a new data point's k nearest neighbours in the features space to classify it or predict a target variable.
5. Support Vector Machine (SVM):Suitable for both classification and regression issues, a Support Vector Machine is a strong and adaptable machine learning technique. Finding a hyperplane in a high-dimensional space that optimally divides the data into distinct classes while optimizing the margin between the classes is the main goal of support vector machines (SVM).
6. Ada(Adaptive) Boosting: Ada-Boost is an ensemble learning technique that combines the results of several weak classifiers to create a strong classifier. Ada-Boost's ability to adaptively modify the weights of incorrectly categorized cases during training is reflected in the term "adaptive" in the system.
7. Gradient Boosting: Gradient Boosting is an ensemble learning technique that builds a strong predictive model by aggregating the predictions of several weak learners, typically decision trees. It is a member of the boosting algorithm family, which trains new models successively to fix mistakes in earlier models.



7.1 Base Learner:

An elementary weak learner, such as straightforward decision tree, is used to begin the process. To create predictions, the training data is fitted to this weak learner.

7.2 Residual caluculation:

The residuals are the disparities between the actual values and the predictions that the present model makes. This is the second phase in the process. These residuals, when applied to classification, show the mistakes made in class label prediction

7.3 Gradient Decent Optimization:

These residuals are educated to be predicted by the weak learners that come after (decision trees). The new weak learner is trained on the residuals of the prior model rather than being fitted to the original target variable. In later iterations, the residuals are supposed to be progressively reduced.

7.4 Learning Rate:

A learning rate parameter is introduced to govern the contribution of each weak learner to the overall ensemble. It makes it possible for the model to be updated more gradually by scaling the contribution of each weak learner. The model is more robust at a lower learning rate, but it takes longer for it to converge.

7.5 Ensemble Combination:

The final ensemble prediction is the result of combining the predictions made by each weak learner. When it comes to classification, this usually entails tallying or voting on each poor learner's guesses.

7.6 Regularization:

Gradient Regularization techniques are commonly used in boosting models to avoid overfitting.

Advantages:

1. High predictive accuracy
2. Handles nonlinear relationships
3. Robust to outliers

DATASET:

A comprehensive dataset for diabetes prediction would include individual records with various relevant features. Common features include:

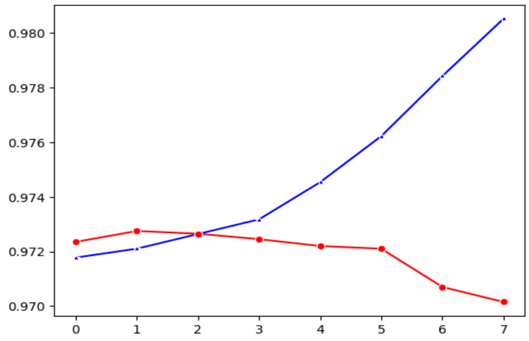
- **Gender:** Gender might be represented as a binary variable, with values such as 0 for male and 1 for female. In some datasets, there might be an additional category for on binary or other gender identities.
- **Age:** Age of the individual.
- **Hypertension:** Hypertension,or high blood pressure, is a medical condition characterized by consistently elevated pressure in the arteries, increasing the risk of heart disease, stroke, and other health complications if untreated.
- **Heart disease:** Heart disease encompasses conditions affecting the heart's structure and function, leading to symptoms like chest pain and shortness of breath, with risk factors including high blood pressure, high cholesterol, and lifestyle habits.
- **Smoking history:** A smoking history in a dataset typically refers to the smoking habits or history of individuals included in the dataset.
- **BMI (Body mass index):** Calculated from height and weight, indicating body fat.
- **HbA1c_level:** refers to the level of hemoglobin in a person's blood. hemoglobin levels may be included as a predictor variable to assess its relationship with diabetes risk or to evaluate its role in disease progression and management.
- **Glucose levels:** Fasting blood sugar levels.
- **Diabetes:** Diabetes is a chronic condition characterized by elevated blood sugar levels due to insufficient insulin production or ineffective use of insulin by the body.

Results:

PARAMETERS	GBC
ACCURACY	0.9755
PRECISION	0.988983
RECALL	0.685261
F1_SCORE	0.809573

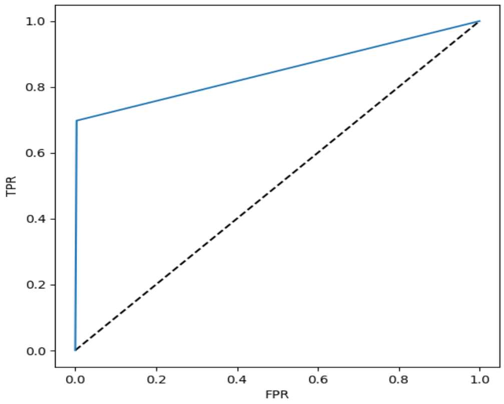
We improved the gradient boosting classifier's accuracy by comparing it to other projects and different foundational articles.

Distribution Curve:



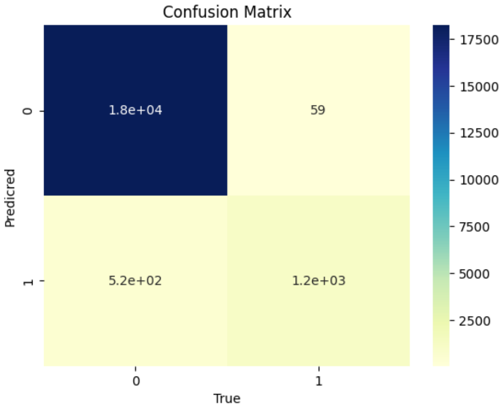
- This graph indicates colours for data distribution, the data is distributed for training and testing.
- For training data it is indicated with blue colour and for testing data it is indicated with red colour.

ROC Curve:



This fig indicates ROC curve for TRP (True Positive Rate) and FPR (False Positive Rate) for predicted and tested values.

Confusion Matrix :



A confusion matrix is a table that is often used to evaluate the performance of a classification model. It compares the actual labels of a dataset with the predicted labels generated by the model. The matrix is typically organized into four sections:
True Positive (TP): Instances where the model correctly predicts the positive class.
False Positive (FP): Instances where the model incorrectly predicts the positive class (false alarm or Type I error).
True Negative (TN): Instances where the model correctly predicts the negative class.
False Negative (FN): Instances where the model incorrectly predicts the negative class (miss or Type II error).
These metrics can provide insights into the model's ability to correctly classify instances of each class and help identify areas for improvement.

Conclusion :

In conclusion, when developing a diabetes prediction system using machine learning techniques, employing the Gradient Boosting Classifier can lead to achieving the best accuracy. Gradient Boosting is a powerful ensemble learning method that combines the strengths of multiple weak learners (decision trees in this case) to improve predictive performance. By leveraging Gradient Boosting Classifier along with appropriate data preprocessing, feature engineering, hyper parameter tuning, and model evaluation techniques, we can build a highly accurate and reliable diabetes prediction model.

REFERENCES:

1. Rashi Rastogi, Mamta Bansal, Diabetes prediction Model Using Data MiningTechniques Volume 25, february 2023,100605.[1]
2. Mitushi Sony, Dr. Sunitha Varma, Diabetes prediction Using Machine learning techniques vol.9, Issue 09,September 2020.[2]
3. Kiran Kumar Patro,Jaya Prakash Allam,Umamaheswararao Sanapala1,Chaitanya Kumar Marpu, Nagwan Abdel Samee,Maali Alabdulhafth and Pawel Plawiak. Patro et al.BMC Bioinformatics (2023)24:372,https://doi.org/10.1186/s 12859 023 05488 6.[3]
4. Isfafuzzaman Tasin,Tansin Ullah Nabil,sanjida Islam,and Riasat Khan,Diabetes prediction using machine learning and explainable AI Techniques 2022 Dec 14[4].
5. KM Jyoti Rani,Diabetes prediction using machine learning ,Volume 6,Issue 4,30 July 2020[5]
6. Ramesh S, Balaji H, Iyengar N, Caytiles RD. Optimal predictive analytics of PIMA diabetics usingdeep learning. Int J Database Theory Appl. 2017;10(9):47-62.
7. Kandhasamy JP, Balamurali S. Performance analysis of classifer models to predict diabetes mellitus. Proc Comput Sci. 2015;47:45–51.
8. Yuvaraj N, SriPreethaa K. Diabetes prediction in healthcare systems using machine learning algorithms on hadoop cluster. Clust Comput. 2019;22(1):1–9
9. Sisodia D, Sisodia DS. Prediction of diabetes using classifcation algorithms. Proc Comput Sci. 2018;132:1578–85.
10. Olaniyi EO, Adnan K. Onset diabetes diagnosis using artifcial neural network. Int J Sci Eng Res. 2014;5(10):754–9.
11. Swapna G, Kp S, Vinayakumar R. Automated detection of diabetes using CNN and CNN LSTM network and heart rate signals. Proc Comput Sci. 2018;132:1253–62.
12. Yahyaoui A, Jamil A, Rasheed J, Yesiltepe M. A decision support system for diabetes prediction using machine learning and deep learning techniques. In: 2019 1st International informatics and software engineering conference (UBMYK. 58 Abdulhadi N, Al Mousa A.
13. Diabetes detection using machine learning classifcation methods. In: 2021 international conference on information technology (ICIT)
14. Abdollahi J, Nouri Moghaddam B. Hybrid stacked ensemble combined with genetic algorithmsfor diabetes prediction. Iran J Comput Sci. 2022;5(3):205–20.