Different Supervised Machine Learning Methods For Predicting And Analyzing the Diabetes

Neenu Augustine
Department of Computer Applications
Amal Jyothi College of Engineering,
Kanjirappally
Kottayam, Kerala.
neenuaugustine@mca.ajce.in

Lisha Varghese
Asst. Professor in Computer
Science Amal Jyothi College of
Engineering, Kanjirappally
Kottayam, Kerala
lishavarghese@amaljyothi.ac.in

Abstract— Diabetes Mellitus is one of the critical essential illnesses that influence masses of humans. Diabetes Mellitus affects how our body uses blood sugar(glucose). It is caused due unregular exercise, our lifestyle, our food habits, high blood pressure level, etc. Many diabetes people have a high risk of different types of diseases such as kidney problems, heart problems, stroke, etc. In the existing system, classification and prediction accuracy is not at a high level. In this paper, Glucose, BMI, blood pressure, Age, etc are the external factors for the prediction of better classification of diabetes. This is the classification technique of supervised system getting to know. This is used to predict whether the patient is diabetes or non- diabetes . There are several predictor variables and one target variable referred to as outcome on the diagnostic measurement on the given dataset. This paper discusses the various applications for predicting and understanding diabetes among people.

Keywords— Supervised Machine Learning, Random Forest, Decision Tree, eXtreme Gradient Boosting, Support Vector Machine.

I. INTRODUCTION

Diabetes is one in every of the harmful sickness in the global.It causes damage to human organs. It mainly affects damage to large blood vessels which can lead to the heart, kidneys, eyes, feet, and nerves. Nowadays many people become diabetes due to their living styles and food habits. . rapid meals is one of the elements causing diabetes. There are two kinds of diabetes as type 1 Diabetes and type 2 Diabetes. kind 1 Diabetes additionally referred to as Insulin-based diabetes used to be referred to as juvenile-onset diabetes which starts off evolved in early life.It affects the body attacks pancreas with its antibodies of body. Humans with kind 1 Diabetes have excessive risks of heart problems and stroke. Type 2 Diabetes is non-insulin-dependent diabetes. It can be seen most commonly in children and teenagers over the past 20 years. .In this form of diabetes, the pancreas creates insulin. in this type, the frame does no longer make sufficient insulin to blood cells and does no longer reply generally to the insulin. Gestational diabetes is another form of Diabetes that develops in some girls throughout the pregnancy stage and it is going away after pregnancy.

II. MACHINE LEARNING

Algorithms or Techiques that enables machines or computers to learns data.A machine learns whenever it is able to utilize its an experience such that its performance improves on similar experiences in the future. Its learning algorithms which can be the mind behinds any of them allows machines to research and less complicated to cause them to. It is a method for Artificial Intelligence systems that conducts the tasks and predicts output values from given input data. There are especially types of strategies along with classification and regression. There are two Kinds of Algorithms, They're of:

- Supervised Learning
- Unsupervised Learning

Supervised Learning is that the method of coaching a prophetic model. It is one of the important algorithms which attempts to optimize a function or model to find the combination of feature values that result in the target output. It is where you have input variables(X) and an output variable(Y) and you use an algorithm to learn the mapping function from the input to the output. The often-used supervised machine learning task of predicting which category an example belongs to its classification.

Unsupervised Learning is a type of Machine learning algorithm in which models were models are not supervised using the training data. It has input data(X) and no corresponding output variables. It is used to model the underlying structure or distribution in the data in order to learn more about data. It can be trained using an unlabeled dataset and performs without any supervision.

III. PROPOSED METHODS

The Goal of this paper used to investigate to predict diabetes with which algorithm gives best accuracy.There are extraordinary varieties of supervised system gaining knowledge of algorithms which include Random Forest, Support Vector Machine, eXtreme Gradient Boosting, decision Tree used for detecting the diabetes. To are expecting the diabetes ,first we collects the diabetes dataset after which pre processing these datas.. Use different types of supervised learing algorithme to detects and predicts the diabetes with high accuracy .

The seven Steps as follows as:

a. Collecting the datas
b. Preparing the datas
c. Choosing the model
d. Training the model
e. Comparing and Evaluating the Model
f. Making the prediction

DOI: 10.5281/zenodo.6369551
ISBN: 978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

The steps that needs for algorithms are of:

a. Import the required libraries.
b. Import the dataset.
c. Define the dataset.
d. Dataset traning and testing.
e. Algorithm Execution
f. Comparison and Evaluation of Results.

Table 1: Dataset Description

SL.NO	Attributes
1	Pregnancy
2	Glucose
3	Blood Pressure
4	Skin Thickness
5	Insulin
6	BMI
7	Diabetes Pedigree Function
8	Age

Figure 1: Description about dataset

A. Data Exploration And Visualizaton

Here, We creates Graphs for displaying different types of the distributions of data and available relationships to understand them and determines how to built the model.

a. Checking Of the distribution of the target variable

From the plot,we can determines that the data with more case are of without diabetes that indicates the value 0 and the data with less case are of diabetes which indicates the value as 1.

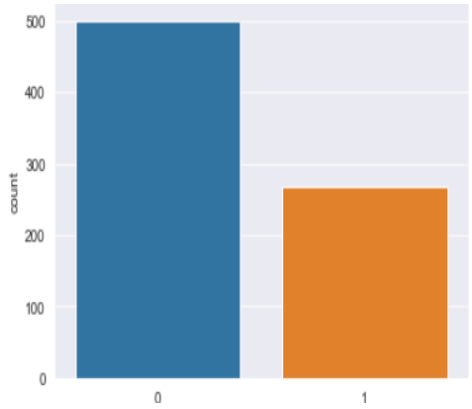

Figure 2:Ratio between Diabetes and Non-diabetes

b. Checking Of the distribution of the predictor variable

Here,we will use plots for every variable to shows its distribution within the dataset.

c. Checking for any lacking values that looks inside the dataset

There are not any missing values inside the dataset and it has been already cleaned.

d. Plotting relationships among the datasets.

There are different ways to display the relationships among the dataset.We can proceed in checking the relationships by visualizing correlations .We can plot the correlations using the heatmap . This is the best way to display the statistical records.

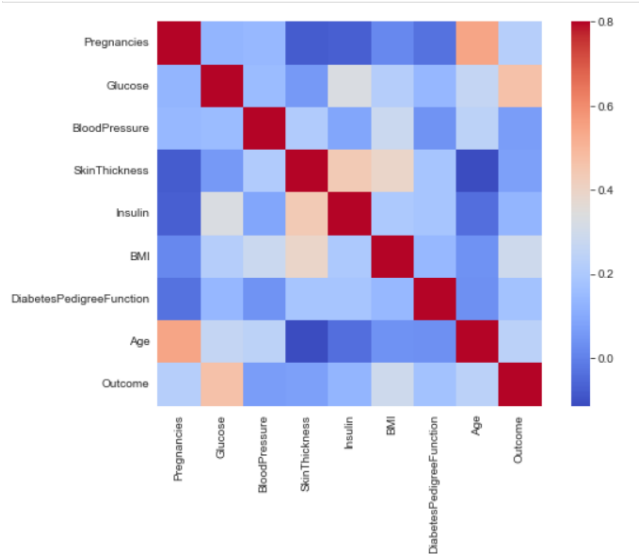

Figure 3. correlation matrix of dataset

B. Training The Data

We split the dataset before trains it. X contains all the Independent variables and Y contains all the Dependent variables.After completing the splitting method it can be slit its train_test_split.Before build the model,we impute the zero values in the dataset.If its contains head of the dataset,its implies that there is a some independent variables with zero values.

C. Apply the Machine Learning Algorithms

There are different types algorithms to predicts the diabetes.The method which is applied based on the diabetes dataset. he algorithms which might be used for predicting the diabetes and their accuracy values are of.They are follow:

a. Random Forest:

It is one of the ensemble studying method used for classification of records and regression tasks.This can be without problems to manages a massive set of datas.It is one of the supervised maching learning algorithms.The greater number of trees that represents highest accuaracy value and prevents its overfitting problems.It also combines for multiple trees for predicting the dataset class.It proiveds output with highest accuracy and maintain them when large set of set is missing.

Algorithms-

- First we pick R features from total features of m,in which R<<m.
- Using best split pointpoint,we select nodes.
- Using split ,we separate the nodes into subnodes.

DOI: 10.5281/zenodo.6369551
ISBN: 978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

- Repeat those steps till of 1 wide variety of nodes which has been reached.
- Repeating theses steps for a number of instances to create n variety of trees for constructing the Forest.

b. Decision Tree:

It is one of the supervised getting to know technique for the fundamental class method for machine learning.it's far one of the tool for classifying and predicting of statistics.It's far a tree-like structure wherein every internal nodes represents a check on attributes related branches denotes an final results and each leaf node that holds a class label.Algorithms-

- Create the tree like structure with nodes with deotes as input feature.
- Select the feature that used to predict the outcome from input feature.
- Repeat these steps to form a subtree using the feature.

c. eXtreme Gradient Boosting:

It is the implementation of the gradient boosting algorithm and its ensembles for type and regression of data.It is the most powerful tool for predicton of data.It is used for improvement of iterations.

Algorithms-

- Consider the sample target values which denotes as P.
- Estimate the error in that target values.
- To lessen the weights,update and adjust the weights
- Those model novices are analyzed and calculated by using loss feature F.
- Repeat these steps desired and target result of P.

d. Support Vector Machine:

It is one of the supervised gadget studying algorithms for category and regression issues.It is used for creating the satisfactory line that can be segregate n-dimensional area to lessons.The hyperplane is the best decision in the SVM.It chooses the vectors for creating the hyperplane.

Algorithm-

- Divides the class by select the hyperplane.
- Calculate the distance between planes and the data (Margins) for finding the better hyper plain.
- If the distance between classes is low then the chances for miss conception is high.
- Then select the class with high margins.

D. APPLY PERFORMANCE MATRIX

Here,We apply different performance matrix such as precision,recall,f1-score and support and confusion matrix.

1. Random Forest

```
[[133  29]
 [ 33  59]]
              precision    recall  f1-score   support

           0       0.80      0.82      0.81       162
           1       0.67      0.64      0.66        92

    accuracy                           0.76       254
   macro avg       0.74      0.73      0.73       254
weighted avg       0.75      0.76      0.75       254
```

Figue-4 Description of various performance matrix in Random Forest

2. Decision Tree

```
[[125  37]
 [ 30  62]]
              precision    recall  f1-score   support

           0       0.81      0.77      0.79       162
           1       0.63      0.67      0.65        92

    accuracy                           0.74       254
   macro avg       0.72      0.72      0.72       254
weighted avg       0.74      0.74      0.74       254
```

Figure-5 Description of various performance matrix in Decision Tree

3. XGBoot

```
[[127  35]
 [ 31  61]]
              precision    recall  f1-score   support

           0       0.80      0.78      0.79       162
           1       0.64      0.66      0.65        92

    accuracy                           0.74       254
   macro avg       0.72      0.72      0.72       254
weighted avg       0.74      0.74      0.74       254
```

Figure-6 Description of various performance matrix in XGBoot

4.Support Vector Machine

```
[[143  19]
 [ 47  45]]
              precision    recall  f1-score   support

           0       0.75      0.88      0.81       162
           1       0.70      0.49      0.58        92

    accuracy                           0.74       254
   macro avg       0.73      0.69      0.69       254
weighted avg       0.73      0.74      0.73       254
```


```
[[143  19]
 [ 47  45]]
              precision    recall  f1-score   support

           0       0.75      0.88      0.81       162
           1       0.70      0.49      0.58        92

    accuracy                           0.74       254
   macro avg       0.73      0.69      0.69       254
weighted avg       0.73      0.74      0.73       254
```


Figure 7: Description of various performance matrix in support Vector Machin

DOI: 10.5281/zenodo.6369551
ISBN: 978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

IV. BUILDING FOR MODEL

This is used for building the model for predicting the diabetes.Here we implemented different types of algorithms for predicting the diabetes.The steps which is used for implementation in the propsed method for predict diabetes.

- First import the required libraries.
- Then import the dataset of diabetes.
- Removal of missing datas.
- Divides the dataset into coaching dataset and testing dataset.
- Then we apply different algorithms such as Random Forest,Decision Tree,eXtreme Gradient Boosting and Support Vector Machines.
- Build a classifier model which depends on selected algorithms using the training set.
- Evaluate the classifier model based on the test set.

.

V. . EXPERIMENTAL RESULTS

Through this work,we uses several supervised learning algorithms for predicting the diabetes.By evaluating and comparing different algorthims, we get Random Forest has higher accuracy than other algorithms.It can be implemented by python language.Through this work, we comes to a decision that Random forest is the best algorithm for predicting the diabetes.Its Model uses to predict the possibility of persons who having the diabetes which represents the value as 1 and the persons which non-diabetes which represents the value as 0.

Table 2: Accuracy Table

SL.NO	MODEL NAME	ACCURACY
1	SVM	74%
2	Random Forest	75%
3	XGBoost	74%
4	Decision Tree	73%

Figure 8: Description about accuracy

VI. . CONCLUSION

The purpose of this paper is used to analysis and compares the distinct types of supervised learning algorithms and evaluates which algorithm gives higher accuracy for predicting the diabetes.Here , Random Forest has higher accuarcy of 75% which is compared to other algorithms.Finally we select Random Forest model for predicting the possibilites of persons having diabetes or not.We can also see that the patient at index 0 indicates absence of diabetes and patient at index 1 indicates chance of diabetes.

VII. . REFERENCE

[1] Mani Butwall and Shraddha Kumar," A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier", International Journal of Computer Applications, Volume 120 - Number 8,2015.

[2] Dost Muhammad Khan1, Nawaz Mohamudally2, "An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm ", Journal Of Computing, Volume 3, Issue 12 DECEMBER..

[3] Shapiro AM, Lakey JR, Ryan EA,et al. Islet transplantation in seven patients with type 1 diabetes mellitus using a glucocorticoid-free immunosuppressive regimen. N.Engl. J. Med. 343, 230-238(2000).

[4] B.M. Patil, R.C. Joshi and Durga Toshniwal,"Association Rule for Classification of Type-2 Diabetic Patients", ICMLC '10 Proceedings of the 2010 Second International Conference on Machine Learning and Computing, February 09 - 11, 2010.

[5] Priyanka Sonar K.JayaMalini Diabetes Predction using different Machine learning Approaches Proceeding of the Third International Conference on computing Methodologies and Communication (ICCMC 2019) IEEE Xplore Part Number: CFP19K25 ART; ISBN: 978-1-5386-7808-4

[6] https://www.healthline.com/health/diabetes