**Diabetes Prediction using Machine Learning**

**Techniques**

***Abstract*:- Diabetes is an illness caused because of high glucose**

**level in a human body. Diabetes should not be ignored if it is**

**untreated then Diabetes may cause some major issues in a person**

**like: heart related problems, kidney problem, blood pressure,**

**eye damage and it can also affects other organs of human body.**

**Diabetes can be controlled if it is predicted earlier. To achieve**

**this goal this project work we will do early prediction of Diabetes**

**in a human body or a patient for a higher accuracy through**

**applying, Various Machine Learning Techniques. Machine**

**learning techniques Provide better result for prediction by constructing**

**models from datasets collected from patients. In this**

**work we will use Machine Learning Classification and ensemble**

**techniques on a dataset to predict diabetes. Which are K-Nearest**

**Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT),**

**Support Vector Machine (SVM), Gradient Boosting (GB) and**

**Random Forest (RF). The accuracy is different for every model**

**when compared to other models. The Project work gives the**

**accurate or higher accuracy model shows that the model is capable**

**of predicting diabetes effectively. Our Result shows that**

**Random Forest achieved higher accuracy compared to other**

**machine learning techniques.**

***Keywords*: *Diabetes, Machine, Learning, Prediction, Dataset,***

***Ensemble***

I. INTRODUCTION

Diabetes is noxious diseases in the world. Diabetes caused

because of obesity or high blood glucose level, and so

forth. It affects the hormone insulin, resulting in abnormal

metabolism of crabs and improves level of sugar in the

blood. Diabetes occurs when body does not make enough

insulin. According to (WHO) World Health Organization

about 422 million people suffering from diabetes particularly

from low or idle income countries. And this could be

increased to 490 billion up to the year of 2030. However

prevalence of diabetes is found among various Countries

like Canada, China, and India etc. Population of India is

now more than 100 million so the actual number of diabetics

in India is 40 million. Diabetes is major cause of death

in the world. Early prediction of disease like diabetes can

be controlled and save the human life. To accomplish this,

this work explores prediction of diabetes by taking various

attributes related to diabetes disease. For this purpose we

use the Pima Indian Diabetes Dataset, we apply various

Machine Learning classification and ensemble Techniques

to predict diabetes. Machine Learning Is a method that is

used to train computers or machines explicitly. Various

Machine Learning Techniques provide efficient result to

collect Knowledge by building various classification and

ensemble models from collected dataset. Such collected

data can be useful to predict diabetes. Various techniques

of Machine Learning can capable to do prediction, however

it’s tough to choose best technique. Thus for this purpose

we apply popular classification and ensemble methods on

dataset for prediction.

II. LITERATURE REVIEW

K.VijiyaKumar et al. [11] proposed random Forest algorithm

for the Prediction of diabetes develop a system which

can perform early prediction of diabetes for a patient with a

higher accuracy by using Random Forest algorithm in machine

learning technique. The proposed model gives the

best results for diabetic prediction and the result showed

that the prediction system is capable of predicting the diabetes

disease effectively, efficiently and most importantly,

instantly. Nonso Nnamoko et al. [13] presented predicting

diabetes onset: an ensemble supervised learning approach

they used five widely used classifiers are employed for the

ensembles and a meta-classifier is used to aggregate their

outputs. The results are presented and compared with similar

studies that used the same dataset within the literature.

It is shown that by using the proposed method, diabetes

onset prediction can be done with higher accuracy. Tejas

N. Joshi et al. [12] presented Diabetes Prediction Using

Machine Learning Techniques aims to predict diabetes via

three different supervised machine learning methods including:

SVM, Logistic regression, ANN. This project proposes

an effective technique for earlier detection of the

diabetes disease. Deeraj Shetty et al. [15] proposed diabetes

disease prediction using data mining assemble Intelligent

Diabetes Disease Prediction System that gives analysis

of diabetes malady utilizing diabetes patient’s database.

In this system, they propose the use of algorithms like

Bayesian and KNN (K-Nearest Neighbor) to apply on diabetes

patient’s database and analyze them by taking various

attributes of diabetes for prediction of diabetes disease.

Muhammad Azeem Sarwar et al. [10] proposed study on

prediction of diabetes using machine learning algorithms in

healthcare they applied six different machine learning algorithms

Performance and accuracy of the applied algorithms

is discussed and compared. Comparison of the different

machine learning techniques used in this study reveals

which algorithm is best suited for prediction of diabetes.

Diabetes Prediction is becoming the area of interest for

researchers in order to train the program to identify the

patient are diabetic or not by applying proper classifier on

the dataset. Based on previous research work, it has been

observed that the classification process is not much im-

proved. Hence a system is required as Diabetes Prediction

is important area in computers, to handle the issues identified

based on previous research.

III. PROPOSED METHODOLOGY

Goal of the paper is to investigate for model to predict diabetes

with better accuracy. We experimented with different

classification and ensemble algorithms to predict diabetes.

In the following, we briefly discuss the phase.

**A. Dataset Description-** the data is gathered from UCI

repository which is named as Pima Indian Diabetes Dataset.

The dataset have many attributes of 768 patients.

**Table 1: Dataset Description**

**S No. Attributes**

1 Pregnancy

2 Glucose

3 Blood Pressure

4 Skin thickness

5 Insulin

6 BMI(Body Mass Index)

7 Diabetes Pedigree Function

8 Age

The 9th attribute is class variable of each data points. This

class variable shows the outcome 0 and 1 for diabetics

which indicates positive or negative for diabetics.

**Distribution of Diabetic patient-** We made a model to

predict diabetes however the dataset was slightly imbalanced

having around 500 classes labeled as 0 means negative

means no diabetes and 268 labeled as 1 means positive

means diabetic.

Figure 1: Ratio of Diabetic and Non Diabetic Patient

**B. Data Preprocessing-** Data preprocessing is most important

process. Mostly healthcare related data contains

missing vale and other impurities that can cause effectiveness

of data. To improve quality and effectiveness obtained

after mining process, Data preprocessing is done. To use

Machine Learning Techniques on the dataset effectively

this process is essential for accurate result and successful

prediction. For Pima Indian diabetes dataset we need to

perform pre processing in two steps.

**1). Missing Values removal-** Remove all the instances that

have zero (0) as worth. Having zero as worth is not possible.

Therefore this instance is eliminated. Through eliminating

irrelevant features/instances we make feature subset

and this process is called features subset selection, which

reduces diamentonality of data and help to work faster.

**2). Splitting of data-** After cleaning the data, data is normalized

in training and testing the model. When data is

spitted then we train algorithm on the training data set and

keep test data set aside. This training process will produce

the training model based on logic and algorithms and values

of the feature in training data. Basically aim of normalization

is to bring all the attributes under same scale.

**C. Apply Machine Learning-** When data has been ready

we apply Machine Learning Technique. We use different

classification and ensemble techniques, to predict diabetes.

The methods applied on Pima Indians diabetes dataset.

Main objective to apply Machine Learning Techniques to

analyze the performance of these methods and find accuracy

of them, and also been able to figure out the responsible/

important feature which play a major role in prediction.

The Techniques are follows-

**1) Support Vector Machine-** Support Vector Machine

also known as svm is a supervised machine learning algorithm.

Svm is most popular classification technique. Svm

creates a hyperplane that separate two classes. It can create

a hyperplane or set of hyperplane in high dimensional

space. This hyper plane can be used for classification or

regression also. Svm differentiates instances in specific

classes and can also classify the entities which are not supported

by data. Separation is done by through hyperplane

performs the separation to the closest training point of any

class.

**Algorithm-**

• Select the hyper plane which divides the class better.

• To find the better hyper plane you have to calculate

the distance between the planes and the data

which is called Margin.

• If the distance between the classes is low then the

chance of miss conception is high and vice versa.

So we need to

• Select the class which has the high margin.

Margin = distance to positive point + Distance to

negative point.

**2) K-Nearest Neighbor -** KNN is also a supervised machine

learning algorithm. KNN helps to solve both the

classification and regression problems. KNN is lazy prediction

technique.KNN assumes that similar things are near to

each other. Many times data points which are similar are

very near to each other.KNN helps to group new work

based on similarity measure.KNN algorithm record all the

records and classify them according to their similarity

measure. For finding the distance between the points uses

tree like structure. To make a prediction for a new data

point, the algorithm finds the closest data points in the training

data set — it’s nearest neighbors. Here **K**= Number of

nearby neighbors, it’s always a positive integer. Neighbor’s

value is chosen from set of class. Closeness is mainly defined in terms of Euclidean distance. The Euclidean distance

between two points P and Q i.e. P (p1,p2, …. Pn) and

Q (q1, q2,..qn) is defined by the following equation:-

**Algorithm-**

• Take a sample dataset of columns and rows named

as Pima Indian Diabetes data set.

• Take a test dataset of attributes and rows.

• Find the Euclidean distance by the help of formula-

• Then, Decide a random value of K. is the no. of

nearest neighbors

• Then with the help of these minimum distance and

Euclidean distance find out the nth column of

each.

• Find out the same output values.

If the values are same, then the patient is diabetic, otherwise

not.

**3) Decision Tree-** Decision tree is a basic classification

method. It is supervised learning method. Decision tree

used when response variable is categorical. Decision tree

has tree like structure based model which describes classification

process based on input feature. Input variables are

any types like graph, text, discrete, continuous etc. Steps

for Decision Tree **Algorithm-**

• Construct tree with nodes as input feature.

• Select feature to predict the output from input feature

whose information gain is highest.

• The highest information gain is calculated for

each attribute in each node of tree.

• Repeat step 2 to form a subtree using the feature

which is not used in above node.

**4) Logistic Regression-** Logistic regression is also a supervised

learning classification algorithm. It is used to estimate

the probability of a binary response based on one or

more predictors. They can be continuous or discrete. Logistic

regression used when we want to classify or distinguish

some data items into categories.

It classify the data in binary form means only in 0 and 1

which refer case to classify patient that is positive or negative

for diabetes.

Main aim of logistic regression is to best fit which is

responsible for describing the relationship between target

and predictor variable. Logistic regression is a based on

Linear regression model. Logistic regression model uses

sigmoid function to predict probability of positive and negative

class.

Sigmoid function P = 1/1+e - (a+bx) Here P = probability,

a and b = parameter of Model.

**Ensembling-** Ensembling is a machine learning technique

Ensemble means using multiple learning algorithms together

for some task. It provides better prediction than any

other individual model that’s why it is used. The main

cause of error is noise bias and variance, ensemble methods

help to reduce or minimize these errors. There are two

popular ensemble methods such as – Bagging, Boosting,

ada-boosting, Gradient boosting, voting, averaging etc.

Here In these work we have used Bagging (Random forest)

and Gradient boosting ensemble methods for predicting

diabetes.

**5) Random Forest –** It is type of ensemble learning method

and also used for classification and regression tasks.

The accuracy it gives is grater then compared to other

models. This method can easily handle large datasets. Random

Forest is developed by Leo Bremen. It is popular ensemble

Learning Method. Random Forest Improve Performance

of Decision Tree by reducing variance. It operates

by constructing a multitude of decision trees at training

time and outputs the class that is the mode of the classes or

classification or mean prediction (regression) of the individual

trees.

**Algorithm-**

• The first step is to select the “R” features from the

total features “m” where R<<M.

• Among the “R” features, the node using the best

split point.

• Split the node into sub nodes using the best split.

• Repeat a to c steps until ”l” number of nodes has

been reached.

• Built forest by repeating steps a to d for “a” number

of times to create “n” number of trees.

The random forest finds the best split using the Gin-Index

Cost Function which is given by:

The first step is to need the take a glance at choices and use

the foundations of each indiscriminately created decision

tree to predict the result and stores the anticipated outcome

at intervals the target place. Secondly, calculate the votes

for each predicted target and ultimately, admit the high

voted predicted target as a result of the ultimate prediction

from the random forest formula. Some of the options of

Random Forest does correct predictions result for a spread

of applications are offered.

**6) Gradient Boosting -** Gradient Boosting is most powerful

ensemble technique used for prediction and it is a classification

technique. It combine week learner together to

make strong learner models for prediction. It uses Decision

Tree model. it classify complex data sets and it is very effective

and popular method. In gradient boosting model

performance improve over iterations.

**Algorithm-**

• Consider a sample of target values as P

• Estimate the error in target values.

• Update and adjust the weights to reduce error M.

• P[x] =p[x] +alpha M[x]

• Model Learners are analyzed and calculated by

loss function F

• Repeat steps till desired & target result P.

Figure 2: Overview of the Process

IV. MODEL BUILDING

This is most important phase which includes model building

for prediction of diabetes. In this we have implemented

various machine learning algorithms which are discussed

above for diabetes prediction.

**Procedure of Proposed Methodology-**

**Step1:** Import required libraries, Import diabetes dataset.

**Step2**: Pre-process data to remove missing data.

**Step3:** Perform percentage split of 80% to divide dataset as

Training set and 20% to Test set.

**Step4:** Select the machine learning algorithm i.e. KNearest

Neighbor**,** Support Vector Machine, Decision Tree,

Logistic regression, Random Forest and Gradient boosting

algorithm.

**Step5:** Build the classifier model for the mentioned machine

learning algorithm based on training set.

**Step6:** Test the Classifier model for the mentioned machine

learning algorithm based on test set.

**Step7:** Perform Comparison Evaluation of the experimental

performance results obtained for each classifier.

**Step8:** After analyzing based on various measures conclude

the best performing algorithm.

V. EXPERIMENTAL RESULTS

In this work different steps were taken. The proposed approach

uses different classification and ensemble methods

and implemented using python. These methods are standard

Machine Learning methods used to obtain the best accuracy

from data. In this work we see that random forest

classifier achieves better compared to others. Overall we

have used best Machine Learning techniques for prediction

and to achieve high performance accuracy. Figure shows

the result of these Machine Learning methods.

Figure3: Accuracy Result of Machine learning methods

Here feature played important role in prediction is presented

for random forest algorithm. The sum of the importance

of each feature playing major role for diabetes have been

plotted, where X-axis represents the importance of each

feature and Y-Axis the names of the features.

Figure 4: Feature Importance Plot for Random Forest

VI. CONCLUSION

The main aim of this project was to design and implement

Diabetes Prediction Using Machine Learning Methods and

Performance Analysis of that methods and it has been

achieved successfully. The proposed approach uses various

classification and ensemble learning method in which

SVM, Knn, Random Forest, Decision Tree, Logistic Regression

and Gradient Boosting classifiers are used. And

77% classification accuracy has been achieved. The Experimental

results can be asst health care to take early prediction

and make early decision to cure diabetes and save humans

life.

VII. REFERENCES

[1] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing

Feature Importance’s for Diabetes Prediction using Machine

Learning". IEEE, pp 942-928, 2018.

[2] K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline,

"Random Forest Algorithm for the Prediction of Diabetes

".Proceeding of International Conference on Systems Computation

Automation and Networking, 2019.

[3] Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance

Analysis of Machine Learning Techniques to Predict

Diabetes Mellitus". International Conference on Electrical,

Computer and Communication Engineering (ECCE), 7-9 February,

2019.

[4] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction

Using Machine Learning Techniques".Int. Journal of Engineering

Research and Application, Vol. 8, Issue 1, (Part -II) January

2018, pp.-09-13

[5] Nonso Nnamoko, Abir Hussain, David England, "Predicting

Diabetes Onset: an Ensemble Supervised Learning Approach ".

IEEE Congress on Evolutionary Computation (CEC), 2018.

[6] Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabetes

Disease Prediction Using Data Mining ".International Conference

on Innovations in Information, Embedded and Communication

Systems (ICIIECS), 2017.

[7] Nahla B., Andrew et al,"Intelligible support vector machines

for diagnosis of diabetes mellitus. Information Technology in

Biomedicine", IEEE Transactions. 14, (July. 2010), 1114-20.

[8] A.K., Dewangan, and P., Agrawal, “Classification of Diabetes

Mellitus Using Machine Learning Techniques,” International

Journal of Engineering and Applied Sciences, vol. 2, 2015.