Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 701 - 711
Received 20 January 2021; Accepted 08 February 2021.

Prediction of Diabetes Mellitus using Ensembled Machine learning Techniques

Oindrila Banerjee*[1],Dr KVV Satyanarayana*[2]

1st author:Oindrilabanerjee(M.TechScholar,KoneruLakshmaiah Education Foundation,Vaddeswaram,A.P,India, Email:oindrubanerjee@gmail.com)

2nd author:Dr KVV Satyanarayana(***Professor, department of CSE &***MTech Coordinator

KL University,email:kopparti@kluniversity.in)

Abstract.

Introduction: Machine learning can be used in the prediction of Diabetes Mellitus thus helping to reduce mortality and morbidity arising out of it. The health system in India is already overburdened and the occurrence of pandemic COVID 19 mandates better predictive outcomes with less interaction in the preliminary stage. Methodology: The study uses 768 records of PIMA Indian Diabetes dataset (500 diabetic and 268 non diabetic records) with 8 attributes of each. The data was pre-processed by replacing the missing values with mean and the performance of the model was increased by removing the imbalance in the dataset by Synthetic Minority Oversampling Technique(SMOTE) algorithm.A new ensembled algorithm SDS (using SGD Classification, decision tree and Gradient Boosting) is presented in the study. Results:The accuracy score of proposed SDS algorithm in the test dataset is 73.38%. and the AUC(Area under curve) value was 70.42%. Conclusion: The proposed ensembled algorithm of Diabetes prediction can be used to support the already overburdened health system.

Keywords: SMOTE,Ensemble technique, SGD classification, decision tree, Gradient Boosting

1. Introduction:

Diabetes Mellitus (DM) can be considered as a health explosion which has taken the world's attention due to the fast spreading tentacles. Approximately 425 million people in the world were diabetic in 2017 and it is expected to reach 629 million by 2045. India has been termed as the diabetic capital of the world due to the highly increasing number of cases and which is projected to increase to 800million by

2030.[1]The southern states including Hyderabad have registered a higher prevalance of Diabetes (16.6%) than the national average of 9.1% in a community based study by Indian Council of Medical Research (ICMR) in 2014.[2]Thepresentera believes in diagnosing diabetes by signs and symptoms, Fasting blood glucose, random blood glucose, oral glucose tolerance test and glycatedhemoglobin. With the advent of technology, various machine learning techniques have been used to predict the diagnosis of DM[3].The Machine Learning technique uses the field of artificial intelligence and is used to devise various models which learns and improves the predictive ability for Diabetes by using prefed data. The technique can not only ease the overburdened patient doctor ratio in our country as a preliminary assessment, it can also serve its purpose in COVID 19 era to minimise human interactions. Such high risk potential diabetic patients can be identified early by machine learning techniques and respective treatment or management can be initiated to prevent the onset or to reduce the complications.

The study proposes a new algorithm which is based on ensembled machine learning technique for prediction of DM. In this study Synthetic Minority Oversampling Technique (SMOTE) algorithm was used for increasing model performance by balancing imbalance data in our PIMA Indian Diabetes dataset.

Aims and Objectives:

a) To predict the occurrence of diabetes mellitus by using machine learning algorithms- Decision tree, SGD classifier, Gradient boosting and ensembled technique (SDS)
b) To evaluate the accuracy in prediction of DM by the machine learning algorithms used

This study is organized as follows - in section2, materials and method is discussed where it comprises of an elaborated explanation of ensemble machine learning model used in this paper , section 3 comprises of the analysis of the results obtained for the ensemble machine learning model, and in section 4concludes the article with discussion.

2.Material and methodology:

2.1 Data set description: The study uses PIMA Indian Diabetes Dataset which is freely available in UCI machine learning repository[4]. The dataset contains 768

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 701 - 711
Received 20 January 2021; Accepted 08 February 2021.

records (268 as diabetic and 500 as non diabetic patients) where each record has 8 attributes and labels to predict DM.They are labelled as 1 and 0 in the class attribute.

Table 1 represents the various attributes taken under consideration.

Table1: Attribute Description:

Attribute	**Description**
Pregnancies	No of times pregnancies
Glucose	Plasma glucose concentration with a 2 hours in an oral GTT
Blood Pressure	Diastolic blood pressure (mm Hg).
Skin thickness	Triceps skin fold thickness in mm.
Insulin	Triceps skin fold thickness in mm.
BMI	Body mass index (weight in kg/ (height in m) ^2).
Diabetes Pedigree Function	Diabetes pedigree function.
Age	Age calculated in years.
outcome	Class variable (0 or 1)

The attribute Diabetes Pedigree Function provides some data about the diabetes history among the patient's relatives and the genetic relationship between that relative and the patient. This attribute is used to get an insight of the patient's hereditary risk regarding the inception of diabetes.

2.2 Data Preprocessing: The best part of the working dataset is that it has no null values in entire dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               768 non-null    int64  
 1   Glucose                   768 non-null    int64  
 2   BloodPressure             768 non-null    int64  
 3   SkinThickness             768 non-null    int64  
 4   Insulin                   768 non-null    int64  
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64  
 8   Outcome                   768 non-null    int64  
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 701 - 711
Received 20 January 2021; Accepted 08 February 2021.

But when the dataset was analysed with python, some of the values of the attributes glucose, blood pressure, BMI was found to be 0 which is practically not possible. Hence imputation was thought to be necessary for a good performance model. Imputation is a process where missing values are replaced by attribute's statistical value. In the study, median was used to replace the missing value. Standard scaler was used for rescaling the dataset in data standardization. Data standardization is a process where one or more attribute of the dataset is rescaled so that they have a mean value of 0 and a standard deviation of 1.

An imbalanced dataset (500 diabetics and 268 non diabetics) can reduce the performance of a model, hence a dataaugmentation technique for the minority class was used and is referred to as the Synthetic Minority Oversampling Technique popularly termed asSMOTE. It works to simply duplicate examples in the minority class, although these examples don't add any new information to the model. Instead, from the existing examplesnew examples can be synthesized. In our study dataset was divided into training and testing parts where partitions is as follows 80% for training and 20% for testing. Since our dataset has only 768 data records so removing a part of it for validation could have posed a problem of under fitting. And also by reducing the training data, there can be a risk of losing important trends /patterns in data set, which in turn increases error induced by bias .So we used K Fold cross validation where the data is divided into k subsets and each time, one of the k subsets is used as the validation set/test set and the other k-1 subsets are put together to form a training set. In our study 10-cross validations are used.

2.3 Ensembled Technique: This technique works by constructing a huge number of classifiers at training time [5].Ensembledmachine learning technique combines several base models in order to produce one optimal predictive model. This technique is a enhancing technique applied to the results of different algorithms to acquire better accuracy that leads to a better classification than individual classifier[6].
Three most popular ensembled techniques are Bagging, Boosting and Voting.

2.3.1 Bagging: It is a process that often considers homogeneous weak learners, learns independently from each of them and other in parallel and combines them following some kind of deterministic averaging process. Different decisions that are normally

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 701 - 711
Received 20 January 2021; Accepted 08 February 2021.

taken from different learners canbe combined into one prediction only[7]. The Bagging algorithm is shown in Algorithm 1.[7].

Input: *dataset* $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$;
Base learning algorithm £;
Number of base learners T.
Process:
(1) *for* $t = 1, \ldots, T$;
(2) $h_t = \mathfrak{L}(D, D_{bs})$ % D_{bs} *is the bootstrap distribution*
(3) End
Output: $H(x) = \arg_{y \in Y} \max \Sigma_{t=1}^{T} \mathbb{I}(h_t(x) = y)$

ALGORITHM 1: Bagging algorithm.

2.3.2Boosting:

This technique attempts to build a strong classifier from a number of weak classifiers. Firstly, a model is built from the training data. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models are added. The Boosting algorithm is shown in Algorithm 2.[7].

Input: *sample distribution* D;
Base learning algorithm £;
Number of base learners T.
Process:
(1) $D_1 = D$. % *Initialize distribution*
(2) *for* $t = 1, \ldots, T$:
(3) $h_t = \mathfrak{L}(D_t)$; % *Train a weak learner from distribution* D_t
(4) $\epsilon_t = P_{x \sim D_t} h_t((x) \neq f(x))$; % *Evaluate the error of* h_t
(5) $D_{t+1} = Adjust\_Distribution\ (D_t, \epsilon_t)$
(6) end
Output: $H(x) = Combine\_Outputs\ (\{h_1(x), \ldots, h_t(x)\})$

ALGORITHM 2: Boosting algorithm.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 701 - 711
Received 20 January 2021; Accepted 08 February 2021.

2.3.3Voting: It is a process which is used in regression or classification problems. It works by creating two or more sub-models and each sub-model makes predictions which are combined in some way, such as by taking the mean or the mode of the predictions, allowing each sub-model to vote on what the outcome should be[8].

2.4Algorithm used for my study:

Decision Tree,SGD and Gradiant boosting classifier .voting classifier was used as ensembled technique in the present study.

A brief graphical representation ofthe proposed algorithm is given in fig3.

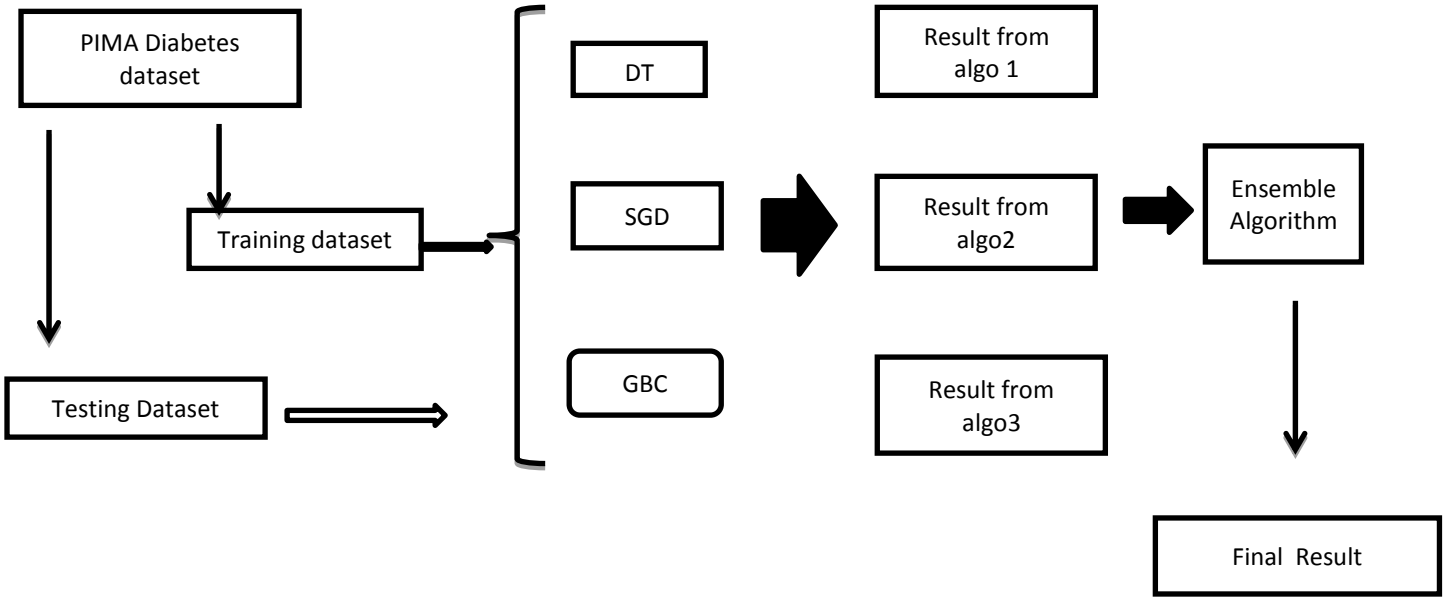
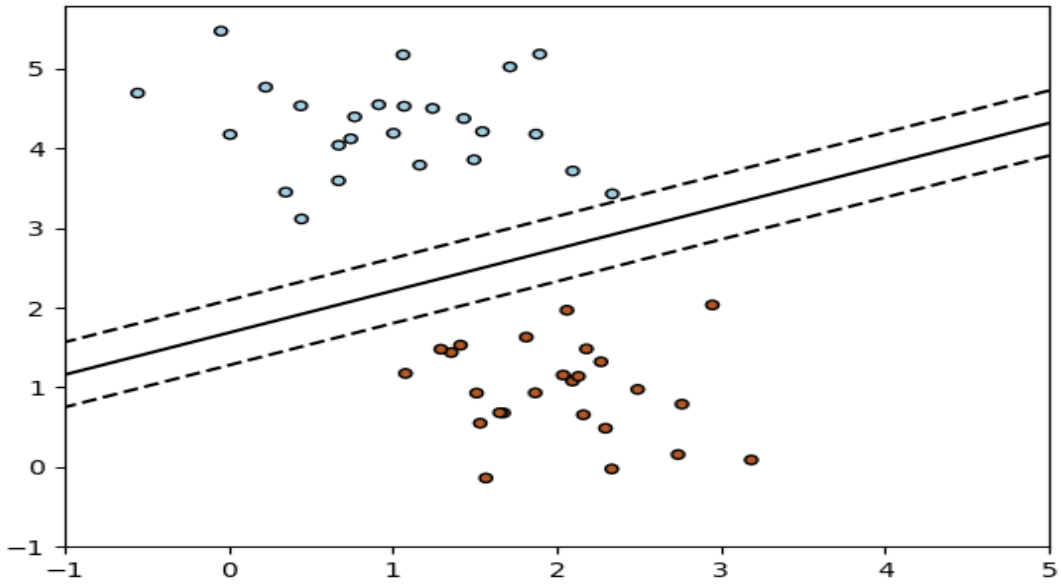


Fig3: Graphical Representation of my proposed algorithm

2.4.1SGD Classifier: SGD stands for Stochastic Gradient Descent approach is used to fit linear classifiers and regressors under convex loss functions such as (linear) Logistic Regression and SVM or Support Vector Machines. A simple and plain stochastic gradient descent learning routine is implemented by SGDClassifier which supports different loss functions and penalties for classification. Below is the decision boundary of a SGDClassifier trained with the hinge loss, equivalent to a linear SVM.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 701 - 711
Received 20 January 2021; Accepted 08 February 2021.



2.4.2Decision Tree: Decision Tree (DT) is a supervised non-parametric machine learning algorithm used for both classification and regression task. A DT's structure is like a flowchart where all internal node are used to represents a test on a future, each and every leaf node are used to represents a class label and all branching in DT represent conjunctions of features that lead to those class labels and root to leaf path represents classification rule. DT can also be considered as a set of if-then rules, which is thought of as a distributions of conditional probability that defined in feature space and class space[9].Ordering attributes as an internal node or root node is a major task and for this task normally statistical method is used.The information Gain can be defined as :

IG(Y,X)=E(Y)-E(Y|X)

Where IG is information gain, and above formula depict information gain from X on Y. here E() represents entropy which is a gauge to measure the level of disorder in data.

2.4.3 Stochastic Gradient Boosting: The statistical framework cast gradient boosting is used as a numerical optimization problem where the objective is to minimize the model's loss by adding weak learners using a gradient descent like procedure. This class of algorithms were described as a stage-wise additive model. This is because one new weak learner is added at a time and existing weak learners in the model are frozen and left unchanged.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 701 - 711
Received 20 January 2021; Accepted 08 February 2021.

3.Result: Three machine learning models were tested in cross-validation set along with theadjusted hyper-parameters for each model.The results show that these models had similar accuracy performance. Hence the models were integrated All to build our final decision model.

Table2:The performance of each model in accuracy

Model name	**Accuracy**
Decision Tree(DT)	68.83%
SGD Classifier	69.48%
Stochastic Gradient Boosting	70.78%

3.1 Combination of Three Models :

After observing the performance of the following three machine learning models Decision Tree,SGDClassifier,andGradiantBoosting Classifier,all the three models were used to generate the final decision model, considering they have very similar performance in CV data set.The voting ensembledtechnique of these three machine learning models were then used to get a better accuracy.

3.2 Evaluation of Model:

The Proposed Model SDS in this study was trained by training data set, and for the performance evaluation of the ensemble model test data was used.Accuracy score,Area under curve (AUC) was used to evaluate the new technique. Where Accuracy score is a common evaluation metric for classification problems. It is a ratio of the number of correct predictions made and total number of predictions made. So,one way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example. The accuracy score of our Proposed model SDS model in the test dataset is 73.38%. and the AUC value was 70.42% asshown in the ROC curve in fig4.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 701 - 711
Received 20 January 2021; Accepted 08 February 2021.

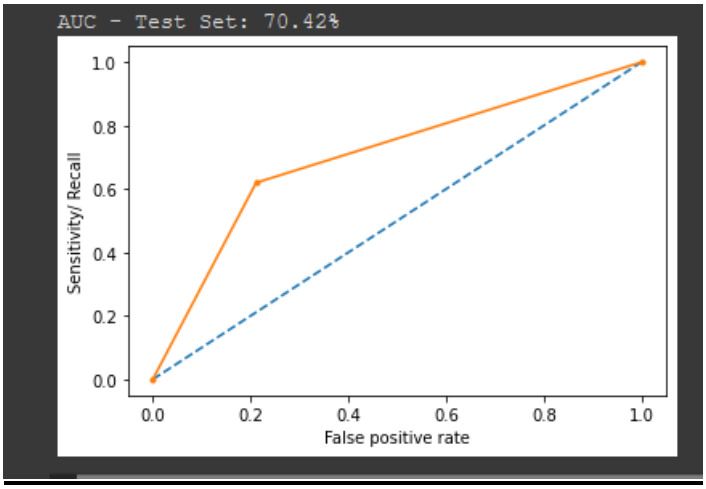


fig4:ROC curve of SDS model

3.3 Results comparison: As its shown in table 2 the performance of the individual model on doing prediction in the measure of accuracy are almost nearby to each other that is 68.83%,69.48%,70.78%where as our proposed ensembled model SDS has the accuracy score of 73.38%.

4.Discussion:

Diabetes Mellitus accounts for more than 90% of patients with diabetes and that leads to microvascular and macrovascular complications that cause keen psychological and physical distress to both patients and carriers and put a massive burden on health-care systems. However, it tends to go undiagnosed as a result of a lack of specific symptoms and limited interest in the public health care sector. Early diagnosis and detection might help in preventing its complications[10] and delaying its progression. Increasing use of Machine learning models in the field of medical science makes it a prominent area of research[11-15]. Machine learning models can be defined as a process to design a model that is learned through experience and to improve its performance.

In this study prediction of diabetes Mellitus has been accomplished using our proposed ensembled model SDS from the PIMA Indian dataset. The dataset quality was improved by the proposed preprocessing scheme, where filling missing values was a core concern. After preprocessing,SMOTE was used to aligh the skewness of attribute's distribution in PIMA indian dataset. The AUC as a weight to build a generic ensemble classifier is better, as it considers more priority to the model having

more AUC.The comparative results reveals that our proposed framework SDS has the higher accuracy score than the individual models that show great potentiality for diabetes prediction from the PID dataset. In the future, a web app with a userfriendly interface can be build with our trained proposed model. Also additionally our SDS model can be applied to other medical contexts to verify their generality and versatility to predict the disease classes.

Limitation of the study:

The study has been done on PIMA Indian Diabetes dataset and can be better validated with the local population data by the methodology used.

Conflicts of Interest

The authors declare no conflicts of interest.

Funding: None

References:--

[1].Patterson CC, Karuranga S, Salpea P, Saeedi P, Dahlquist G, Soltesz G, Ogle GD. Worldwide estimates of incidence, prevalence and mortality of type 1 diabetes in children and adolescents: Results from the International Diabetes Federation Diabetes Atlas. Diabetes research and clinical practice. 2019 Nov 1;157:107842.

[2].Pandey SK, Sharma V. World diabetes day 2018: battling the emerging epidemic of diabetic retinopathy. Indian journal of ophthalmology. 2018 Nov;66(11):1652.

[3].Barber SR, Davies MJ, Khunti K, Gray LJ. Risk assessment tools for detecting those with pre-diabetes: a systematic review. Diabetes research and clinical practice. 2014 Jul 1;105(1):1-3.

[4].UCI (2017) Pima indians diabetes data set. https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes. Accessed 20 Dec2017

[5]. Dietterich TG. Ensemble methods in machine learning.InInternational workshop on multiple classifier systems 2000 Jun 21 (pp. 1-15).Springer, Berlin, Heidelberg.

[6].Sarwar A, Ali M, Manhas J, Sharma V. Diagnosis of diabetes type-II using hybrid machine learning based ensemble model. International Journal of Information Technology. 2020 Jun;12(2):419-28.

[7].Yaman E, Subasi A. Comparison of bagging and boosting ensemble machine learning methods for automated EMG signal classification. BioMed research international. 2019 Oct 31;2019.

[8].Kabari LG, Onwuka UC. Comparison of Bagging and Voting Ensemble Machine Learning Algorithm as a Classifier.International Journals of Advanced Research in Computer Science and Software Engineering. 2019;9(3):19-23.

[9].Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics. 2018 Nov 6;9:515.

[10].Sarwar A, Ali M, Manhas J, Sharma V. Diagnosis of diabetes type-II using hybrid machine learning based ensemble model. International Journal of Information Technology. 2020 Jun;12(2):419-28.

[11]. Ibrahim L, Mesinovic M, Yang KW, Eid MA. Explainable Prediction of Acute Myocardial Infarction Using Machine Learning and Shapley Values. IEEE Access. 2020 Nov 24;8:210410-7.

[12]. Sun ML, Liu Y, Liu G, Cui D, Heidari AA, Jia WY, Ji X, Chen H, Luo Y. Application of Machine Learning to Stomatology: A Comprehensive Review. IEEE Access. 2020 Oct 5;8:184360-74.

[13]. Li JP, Haq AU, Din SU, Khan J, Khan A, Saboor A. Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare. IEEE Access. 2020 Jun 9;8:107562-82.

[14]. Kaur S, Singla J, Nkenyereye L, Jha S, Prashar D, Joshi GP, El-Sappagh S, Islam MS, Islam SR. Medical Diagnostic Systems Using Artificial Intelligence (AI) Algorithms: Principles and Perspectives. IEEE Access. 2020 Dec 3.

[15]. Zhang Z, Han Y. Detection of ovarian tumors in obstetric ultrasound imaging using logistic regression classifier with an advanced machine learning approach. IEEE Access. 2020 Mar 3;8:44999-5008.