PREDICTION OF DIABETIC DISEASE USING ENSEMBLE CLASSIFIER

[1]P.Kalaiyarasi [2] Dr.J.Suguna

Abstract--- *In today's human life, diabetic is the most vulnerable and non-communicable disease creating a great impact in their life. Change in lifestyle and work culture of the people results in millions of diabetic in 21stcentury. Huge amount of data are generated in the modern world, by means of computational analytics on clinical big data. This data are put intocreating a medical intelligence that could be drive the forecasting and prediction. This development in medical intelligence results in great benefit to the people by reducing the hospital re-admission and medical cost, by making this system a patient-centric. Reducing the optimal cost and run time is the result provided by means of improving the health care system by data analytics. In this paper, thediabeticdata aregathered from Kaggle repository. Initially, data has to be pre-processed and randomly divided into training and testing data. Then different ensemble algorithms namely, Bagging, Boosting and Stacking are used for predicting the diabetic disease. Finally ensembleclassifiers results are evaluated by using variousvalidation metrics.*

Keywords--- Big Data Analytics, Diabetic Data, Imbalanced Class, Ensemble Classifiers, AdaBoost.

Introduction :

Today's healthcare industry is facing the most expensive challenge called Diabetic. More than 400 million people among the world are facing this health issue. According to World Health Organization (WHO), 1.5 million deaths are caused directly by Diabetic and high blood glucose resulted in 2.2 million deaths based on 2012 survey. If this situation continues, then Diabetic will affect over 700 million people among world by 2025[24].In recent time'slarge amount of medical data is being generated by healthcare industry.

The information which are created from Electronic Health Records (EHRs) [10] incorporates clinical data and hereditary information of high throughput. Electronic Health Reports (EHR) of specialist's recommendation, drug store data, therapeutic pictures, medical coverage related information, restorative diaries, social medians, patient's information, clinical information and indicative reports are incorporated into the wellbeing care. All these data are collectively shapes a Enormous Information in wellbeing care. The investigation of enormous

[1] Ph.D. (Research scholar), Department of Computer Science, Vellalar College for Women, Tamilnadu, India
[2] Associate professor, Department of Computer Science, Vellalar College for Women, Tamilnadu, India
kalaipavi6@gmail.com, sugunajravi@yahoo.com

information will create the outcomes for understanding the patterns to improving the existence time hope, ease treatment at beginning periods. This examination related with four attributes in particular volume, speed, assortment and veracity [18].

This Big Data analytics mainly focuses on identifying the patients with care gaps and additional support for the patients to reduce the medical expenses. Significant amount of data is produced by means of remarkable advances in health science and biotechnology. Use of information mining and AI techniques to this end is more indispensable and vital efforts for transforming the raw information into useful knowledge. Diabetic disease is an effective metabolic disorder applying significant pressure on human health among the world. Huge amount of data is being generated on performing wide range of research on all aspects of diabetic (Diagnosis, Therapy, etc.) [7].

Diabetic data, because of its unstructured growing nature or all other sources, there is a necessity for structuring and emphasizing the size into ostensible incentive with conceivable arrangement. There is a necessity for combining the electronic communication system and robust diabetic data sharing with the aid of technological developments, which in terms facilitate better service to all patients in terms of both quality and cost. This will bring about all patient data to be put away in a solitary archive.

Clinical information from several repositories are extracted by means of deploying Health Information Exchange(HIE) and combine that data inside a solitary patient wellbeing record which could be gotten to safely by all consideration suppliers. Assortment of procedures namely, statistics, information mining and game hypothesis are being incorporated in predictive analysis for predicting certain future events which employs past and current data with statistical and analytical models. By utilizing the huge information examination in human services, important decisions and prediction could be made in the health care industry on treating the patients with diabetic [18]. Employing big data in management of diabetic in general is a novel approach and new data will aid greatly in better understanding of disease and helps in gaining new knowledge on the disease.

Variables on large amount is being gathered, integrated and examined which would highlightcertain factors with the multifaceted aspects of diabetic management resulting in improvement of interest rate in healthcare community. Collection of observational, “non-experimental” information is processed in big data, which comprises of unclear and potential facts grounded in data [12].Medical intelligence is created by the massive amount of data gathered by means of applying computational analytics on clinical data, which will drive forecasting and medical prediction on the data. The medical intelligence developed with the clinical data will result in a system to be patient-centric thereby reducing the re-admission of patients and also results in reduced cost. [21].

The main focal point of this work is to identify the appropriate machine learning algorithm for diabetic disease prediction. The purpose which falls next in line is providing the overview of benefits on applying the big data analytics to diabetic management by the current researchers. The new horizons and predictions are introduced for diabetic disease management. This paper also focuses on investigation of algorithm in a predictive way for accurate classification. Prediction of diabetic disease is the primary center of this paper. The Kagglerepository is the source for collecting data. The data are preprocessed by SMOTE, a well-known algorithm handles the problem of class imbalance among the classes.The machine learning algorithms namely, AdaBoost, Bagging and Stacking are used for classification.This research mainly focuses on handling theimbalancedclassesby using SMOTE, selecting the suitable machine learning algorithm for classification, and finding the optima classifier providing the more accurate results for the diabetic disease prediction. The remainder of this paper is sorted out as pursues. Section 2 discussed with associated workings of big data health care analytics; Section 3 gives points of interest of the proposed system. In Section 4, test comes about are portrayed. Section 5 gives the final conclusion.

2. BACKGROUND STUDY

Bhavana N, et al. [1] combined the most understood methodologies; Naive Bayes (NBs), K-Nearest Neighbor, J48 and Random Forest (RF) into one as ensemble model. The experiments results reveals that this ensemble model increase the accuracy. The representation serves to be valuable by specialists and pathologist for the reasonable wellbeing the board of diabetes.

Herbert F. et al. [5] compared the effectiveness of Alternating Decision Tree (AD Tree), J48, Naive Bayes Tree (NBTree), Random Tree, Reduced Error Pruning Tree (RepTree) and Simple Cart decision tree classifiers.Likewise examined and looked at the adequacy of AdaBoost, Bagging, MultiBoost, Stacking, Decorate, Dagging, and Grading, in light of Ripple Down Rules as instances of troupe classifiers to characterize the cardiovascular autonomic neuropathy infection. What's more, established that Random Forestperformed best as a base classifier, and AdaBoost, Bagging and Decorate accomplished the best results as meta-classifiers. .IoannisKavakiotiset al. [7] used wide range of AI calculations for the diabetic research. Supervised learning approaches are used by 85% and unsupervised approaches are used by 15% in this system for prediction of diabetes.

Kemal Akyol et al. [9] employed the feature selection methods and broke down to locate the best characteristics required for diabetic illness. The exhibitions of AdaBoost, AdaBoost, Gradient Boosted Trees and Random Forest are assessed. The expectation exactness of the mix of Solidness Choice technique and AdaBoost learning calculation is minimal superior to different calculations with the grouping precision of 73.88%.

NongyaoNai-arun et al. [12] proposed information mining methods to improve productivity and unwavering quality in diabetic disease characterization. The three well understood calculations naïve bayes, k-nearest neighbors and decision tree were utilized to develop arrangement models on the chose highlights. At that point, the gathering learning; stowing and boosting were connected utilizing the three base classifiers. The outcomes uncovered that packing model yields the most noteworthy exactness with base classifier decision tree algorithm of 95.312%. The investigations results demonstrated that gathering classifier models performed superior to the base classifiers alone.

PunneeSittidech et al. [14] united the sacking strategy with base classifier decision tree and cost delicate investigation for diabetic patients' arrangement. The significance elements were chosen and used to develop base classifier choice tree models to group diabetic and non-diabetic patients. The stowing strategy was applied to improve the accuracy. The models are useful for specialists to analyze quiet hospitalization likelihood and to propose some potential medicines to help improve social insurance quality.

Saba Bashir et al. [16] combined three types of decision trees namely, CART, C4.5 and ID3.AdaBoost, Stacking, Majority Voting, Bagging and Bayesian Boosting are taken for ensemble classifier to classify the diabetic disease. The experiments results indicated that bagging ensemble outperforms other techniques for the diabetic classification.

Yukai Li et al. [22] utilized attribute choice and unfair procedure to characterize and predictdiabetic information. After attribute determination and unequal procedure was utilized as info factors of, support vector machine (SVM),), decision tree and coordinated learning mode (Sacking and AdaBoost) for displaying and expectation. The exploratory outcomes demonstrated that AdaBoost calculation creates better order results.

SajidNagiet.al [17] employed the ensemble algorithm namely; Bagging, Boosting and Stack generalization with the base classifiers J48, NB and IBK and used toclassify the micro array cancer data. The experimentresults concludedthat bagging with NB is best for classifying the micro array data. Seokho et al [19] projected a productive and viable troupe of SVM called E-SVM for the counter diabetic medication breakdown forecast. The proposed technique rejects unnecessary information focuses when developing a SVM outfit, there by yielding a superior arrangement execution. Trial results demonstrated that the SVM is reasonable for the counter diabetic medication disappointment forecast with 80% precision.

3. PROPOSED METHODOLOGY

Kaggle repository acts as a source of the proposed system for obtaining thediabetic dataset. The initial step is to pre-process the information so as to balance the classes. SMOTE is employed for balancing the imbalanced classes. The dataset is further classified into training dataand testing data. Later a prediction model is being developed utilizing the test data and the same is evaluated using the testing data. The base classifiers namely Decision Tree, K-Nearest

Neighbor, Support Vector Machine are employed for diabetic disease prediction [8]. In this work, the ensemblelearning algorithms namely, AdaBoost, Bagging and Stacking are used for data classification. The presentation of ensemble learning algorithms are evaluated by means of classification accuracy. The proposed methodology consists of the following stages.

Step1: Input the diabetic dataset.
Step2: Preprocessing the dataset using SMOTE algorithm for balancing the classes.
Step3: The dataset has been randomly split into training and testing data.
Step4: Applying the machine learning algorithms to predicting the diabetic disease.
Step5: Finally the performances of classifiers are evaluated.
The figure 1 represents the design of the proposed framework.

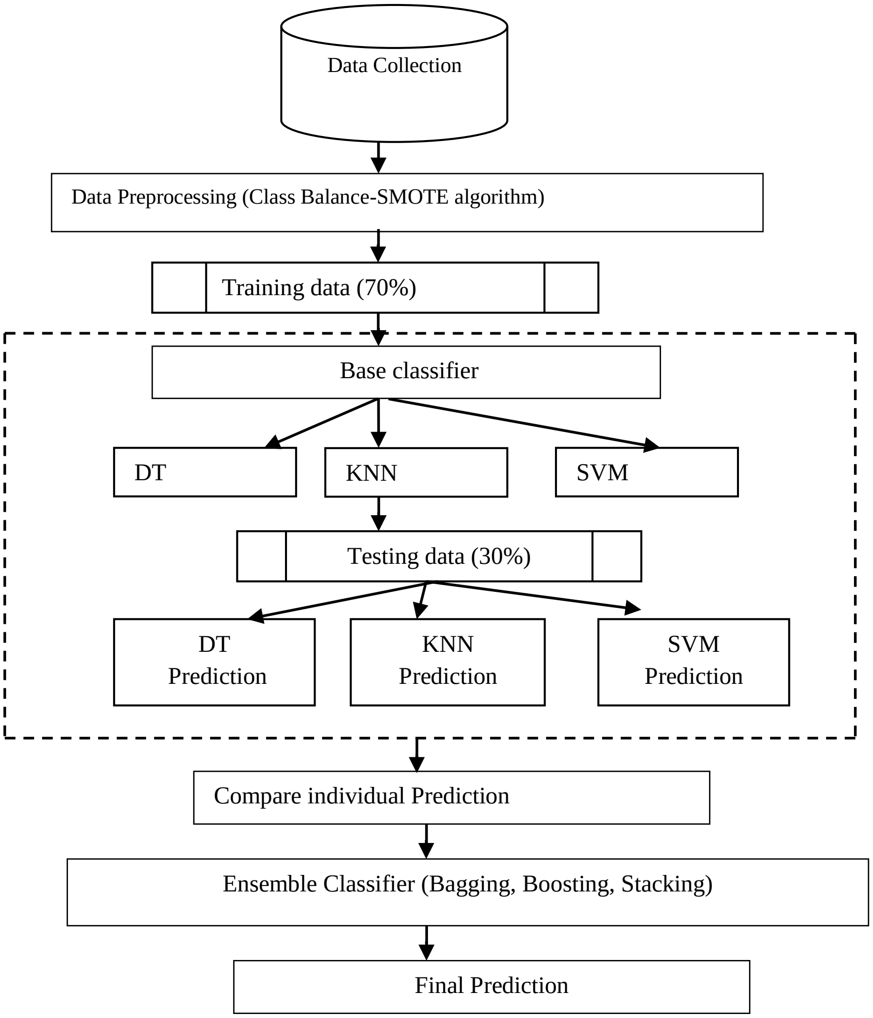

Figure 1: Proposed System Architecture

Data Collection : Kagglerepository [27] is the source from which the dataset is obtained and it is depicted in the following table.

Dataset	Instances	Attributes
Diabetic data	786	9

Table 1: Dataset Description

The diabetic data set contains two classes with 9 attributes such as pregnancies, glucose, blood pressure, Insulin,BMI, diabetes pedigree function and so on.

Preprocessing by SMOTE

Enormous data and its distribution is one of the largest stumbling blocks in any process. Machine learning algorithms at almost time faces the class unevenness problem, where the complete number of positive class of information is far not exactly the all outnumber of negative class of information. The imbalanced dataset present in the data makes the machine learning algorithms to generate unsatisfactory classifiers. The diabetic data poses two classes namely class 0and class1. The number of people with diabetic is very much lesser when compared to number of people with no diabetic. The main focus is improving the rare minority class for achieving the overall accuracy in this prediction.

If the event which is to be predicted belongs to the minority class or if the event rate is lower than that of the threshold value of 5%, thenthe event is considered to be a rare event. The accurate measurement of model performance could not be obtained in the conventional model evaluation methods while facing the imbalanced dataset. Logistic Regression and Decision Tree are the standard classifier algorithms consists of number of instances that have bias towards classes. Majority class data is predicted by them. Minority class comprising features are considered to be noise and often they are ignored. Thus the misclassification of the minority class in comparison with majority class occurs at the high probability. Improvement in the classification algorithm or balancing classes is the strategy which is entailed while dealing with imbalanced classes prior to loading data into machine learning algorithms as input.

Enhancing the frequency of minority classes and reducing the frequency of majority classes is the fundamental target of adjusting classes. By performing these actions, both the classes could acquire around a similar number of examples. In this paper, in order tobalance the classes SMOTE is employed.By taking every minority class test and presenting artificial models along the line fragments joining any of the k minority class closest neighbors the marginal set is over-tested. Contingent on the measure of over-inspecting vital, neighbors from the k closest neighbors are arbitrarily picked.

Bootstrapping and k-nearest neighbors are employed by SMOTE for generating the artificial data. Precisely, it works this way:
Step1: Distinction between the element vector (test) viable and its closest neighbor is taken.
Step2: An rough number somewhere in the range of 0 and 1 is multiplied with this difference.
Step3: Results are added to the attribute vector viable under consideration.
Step4: This creates the choice of an rough point along the line section between two explicit highlights. Step5: Assign the value to the new engineered marginal class test.
Step 6: Repeat the process for identified feature vectors.

BASE CLASSIFIER

Decision Tree

Inside the sort of a tree structure the decision tree constructs classification or regression models. Each leaf hub is relegated a class mark in a decision tree,.the root and other inner hubs is incorporated in the non-terminal hubs, contain credit test conditions to separate records that have various attributes[26].

The feature preference measure is a heuristic for choosing the parting foundation that "best" separate a given information division D, of class marked training tuples into individual classes. Preferably each segment would be clean, if we somehow happened to divide D into littler segments as per the results of the parting criteria. The feature determination measure gives a positioning to each trait portraying the given preparing tuples. The characteristic having best score for the measure is picked as the parting trait for the given tuples. If the parting property is nonstop esteemed or on the off chance that we are limited to binary trees, at that point, individually, either a split point or a parting subset should likewise be resolved as a element of the parting measure. In this work information gain trait choice measure is utilized to choose the root node of the tree. We want to estimate entropy of class label, to build a decision tree.

$E(s) = \sum_{i=1}^{c} -p_i \, log_2 p_i$--------1
where p_i stands for probability of item with the class.
Entropy with twoattributes
$E(T,X) = \sum_{C \in X} \big(P(C)E(C)\big)$------2
where P(C) denotes the percentage of each class, E
(C) stands for entropy of class.

Decision Tree Algorithm Pseudo code
Step1: The best attributes of dataset are set at the root of the tree.
Step2: The guidance set is being divided into partitions. Same value of attributes must be present in a data of each subset.

Step3: Repeat step 1 and step 2 on every division until the leaf nodes are identified in all the branches of the tree.

K-Nearest Neighbor

The Nearest neighbor classifiers depend on learning by relationship, that is, by contrasting a given test tuples and preparing tuples that are like it. KNN has been utilized in factual estimation and pattern acknowledgment as a non – parametric procedure

KNN is considered to be the supervised learning technique in which classififcation of points occur for the given category with the aid of training set.Let (Xi, Ci) where i = 1, 2……., n be data points. Xi stands for feature values &Ci indicates

marks for Xi for every i. The guidance tuples are portrayed by n properties. Each tuple speaks to a point in an n-dimensional space. At the point when given an obscure tuple, a KNN classifier scans the example space for the k guidance tuples that are nearest to unknown tuple. These k guidance tuples are the k "closest neighbors" of the unclear tuples. Nearness is characterized as far as a separation metric, such as Euclidean separation, which is defined as,

(X1, X2) = sqrt (sum ((xj-xij) ^2) ------------3

where, x - new point

xi – existing point across all input attributes j.

KNN AlgorithmPseudo code

Step1: Determine the K training instance which are closest the unknown class instance

Step2: Find the distance between instances and training samples

Steps3: The guidance samples are arranged to the nearest neighbor depends on the distance

Step4:Choosethe most commonly occurring K instance values.

Support Vector Machine

SVMis greatly employed for regression and classification in the medical diagnosis and is considered as a set of interconnected supervised learning models [6].Simultaneously SVM minimizes the empirical classification error and it maximizes geometric margin. Sometimes it is called as Maximum Margin Classifiers. where, to find non-linear classification efficiently we use kernel trick. SVM are represented by some examples, where some separate categories are being used to split by marginal gap which is wide as possible it can. Labeled training data are given as data points:

M= {(x_(1,) y_1), (x_2,y_2)… (x_n,y_n)}------------4

Where yn=1/-1 is constant which denotes the class for the point xnbelongs, n, number of data samples and xn. in general p-dimensional real vector. The classification is performed by SVM by employing appropriatethreshold value, only after SVM classifier maps the input vectors into a decision value. We divide (or separate) the hyperplane for viewing the training data, which is described below:

Mapping: w^T.x+b=0 ------------------5

Where w stands for p-dimensional weight vector,b indicates scalar. Wis the vector which points perpendicular for separating hyperplane. The margin is increased by the offset parameter b. The hyperplane mandatory to pass through the origin in absence of b and restricts the solution.

SVM Algorithm Pseudo code

```
Step1:  Initially identify the right hyperplane
Step2: To maximize the distance between neighbor data points and hyper plane
Step3: Add features to the data points
Step4:  Finding the hyper plane to classify the class.
```

B) ENSEMBLE CLASSIFIER

An ensemble method is used to constructs a pair of base classifiers from training data and it tries to performs classification obtaining the predictions made by each base classifier [26]. An ensemble lean towards to be more accurate than its base classifiers. To buildvariousclassifier from the original data and later tosummative their predictions while classifying the unknown class, the ensemble classifier can be constructed by the following ways:

1) By manipulating the training set:In this method, According to some sampling distribution multiple training sets are formed by resembling the original data and then it classifies from each training set using base learning algorithm.

2) By manipulating the input features: In this advance, a subset of input feature is been chosen to form each training set as well the subset can be selected randomly.

3) By manipulating the class labels:In this approach, we canuse when the number of classes is sufficiently high in numbers. The training data are transformed into a binary classification problem by randomly partitioning the class labels into two disjoint subsets subsequently.

4) By manipulating the learning algorithms: There are lots of learning algorithms which can be manipulated in such a way that, it can apply the algorithm numerous times on the same training data where it results in different models.

The key objective of ensemble method is to advance the performance of the base classifier. In this work, three ensemble algorithms - Bagging, Boosting and Stacking are used for classification.

Bagging

Bagging is one of the technique which is used to repeat sample from dataset as a result to obtain uniformly probability distribution, bagging is sometimes also known as bootstrap aggregating.

Each bootstrap sample is of the similar size of the original data. Sampling are performed by some replacement,there are possibilities where some instances may appear several times in the same training sets, where someothers may be left out from the training sets.Given a set, D, of d tuples, bagging as follow. For iteration i (i=1, 2…, k), a training set, Di, of d tuples is testedthroughadditional from the original set of tuples D. here we consider each training set as a sample bootstrap. The sampling along replacement is used with the original tuples of D, which may not be included in Di, whereas others may follow more than once. A classifier model, Mi, is learned for each training set, Di toclassify an unknown tuple, X is termed as classifier Mi which returns its class prediction and it is counted as one vote. The bagged classifier M*, counts the votes and assigns the class with the most votes to X. Bagging can be applied to the prediction of continuous values by claiming its average value of each prediction for a given testing tuple.

Boosting

Boosting is an iterative bagging process where models are trained in a sequential order. Different base learners are created by boosting in training dataset on sequentially reweighting the instances, unlike bagging. Larger weight will be obtained by the previous base learners while misclassifying each instance while taking a set from training data of the next round. Boosting has always been one of the elementaryidea while frequentlywe try apply for base learners to produce sequence of base learners on the total iterations number which are predefined [13].All instances in general are of uniform weights.

For each boosting iteration after initialization makes fit of base learner. Higher weights will be gained by improperly classified instances where the classified weights instances arelet down while computing error. Linear combination from numerous base learners, considered asa model finally obtained by boosting the algorithm which is weighed by their own performance. In this work, adaptive boosting is employed for classification. To reduce the inherent over fitting problem which is actually present in machine learning is possible via adopting AdaBoost technique, where Adaptive Boosting (AdaBoost) is anensemble classifier method.

$F(x)=sign(\sum_{m=1}^{M}(\theta_m f_m)(x))$ ------------6

Where mthweak classifier is denoted by f_m and corresponding weight is denoted by θ_m. Boosting classifer with three base classifiers are used to build the ensmebel model.

Stacking

The popular ensemble learning is Stacking and high level base learner employs the general methods for achieving high predictive accuracy by combing the lower level base learners. Stacking is similar to boosting. Stacking does not allow the base learners of same type to combine whereas the Boosting and Bagging allows it. The stacking are used and it’s also appliedwith base learners along with different machinelearning algorithms.

The Two phases and its Performance tasks of Stacked Generalization:

(i) Bootstrapped samples taken from the Training dataset of Level -0 to train Base classifiers of layer-1.

(ii) The resultant outputobtained from layer-1 are utilized and proceeded to train meta-classifier from layer-2.

The objective is verify whether, the training data is been learned properly. For instance, if one of the classifier is not learning its certain region correctly of the feature space where it leads toits consistency of misclassifying instance coming from that region, at that time the level-2 classifier will adopt to learn the present behavior along with learned behaviors of the classifiers as well it tries to correct the improper training set.

The baseclassifiers individual predictions are combined as follows:

Step: 1. In diabetic data, if all the base classifiers try to predict the same class, then obtain the same decision which goes by ensemble.

Step: 2.While predicting majority classifiers (2 of 3) matches- then, the below steps areperformed :

(a) Class0 is an expert as it tries to predict from the class. Where, class1 does not try to predict as class 0, if that is the result then the prediction obtained by C0 is considered as the decision of ensemble.

(b) We say class0 as an expert, as it tries to predict anyone from the classifier, we also say class1 to be also an expert in its predictions, where the ensemble looks for the class probabilities of the respective classifiers and it tries to select the highest value. If at all there is any relation between the probability values, then the ensemble goes with its majority.

Ensemble Classifier (Accuracy %)	Base classifiers

Step: 3. While Predicting all classifies if itdisagrees, then the any of the following situations listed below shall arise:

(a) While predicting any one of the classifiers shall be an expert if that is the result then, the ensemble will go by classifiers decision.

(b) There are chances where 2 classifiers shall be experts while predicting a class at that situation, the highest class probability is taken as the final result of the classifier.

(c) There are possibilities where all the above three classifiers shall be experts in class predictions.

If that is the case then, The decision of the classifier is considered by taking the highest class probability as the final decision.

By this way, Class Predictions from base classifiers are combined to obtain final prediction.

4. EXPERIMENTAL ANALYSIS

The SMOTE algorithm greatly eliminates theim balance classes thereby improving the accuracy of the classification. Three classifiers which are Decision Tree, K-Nearest Neighbor and Support Vector machine algorithms are used as base classifiers and three ensemble classifiers namely AdaBoost, Bagging and Stacking are used for data classification. Table 2 gives the confusion matrix.

	DT	KNN	SVM
Bagging	74.48	66.67	70.83
Boosting	71.73	72.17	76.95
Stacking	71.74	72.17.	76.95

Table 2: Confusion Matrix

To know the better Classifiers performance we use Confusion matrix table. To know the accuracy percentage of classification the correct number of ratio is predicted and the total predicted numbersare multiplied by 100. The table generally contains 2 rows and 2 columns for binary classification problem. The class labels are observed across on the top and the predicted labels are observed on the side. Number of predictions made is contained in each cell.

Table 3: Accuracy before Preprocessing

Actual//Predicted	Positive	Negative
Positive	TP	FP
Negative	FN	TN

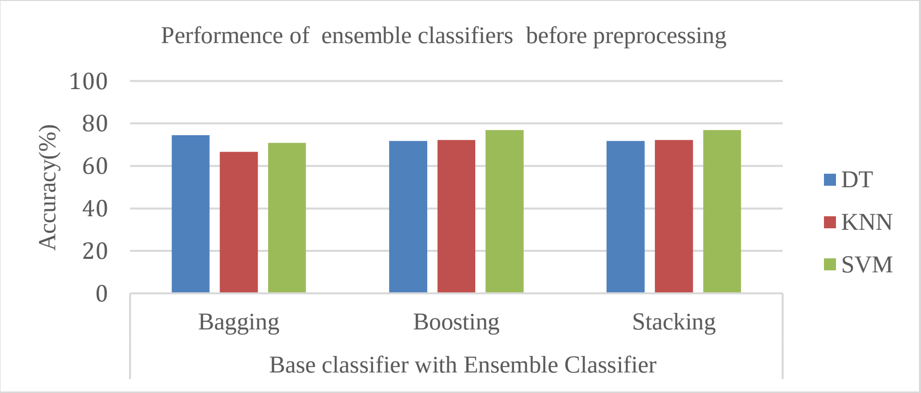

Figure 2: Accuracy before preprocessing.

Validation Measure	Algorithms/ Imbalanced Class								
	Boosting			Bagging			Stacking		
	DT	KNN	SVM	DT	KNN	SVM	DT	KN	SVM
Precision	0.77	0.78	0.78	0.75	0.73	0.77	0.76	0.78	0.78
Recall	0.81	0.81	0.90	0.76	0.73	0.78	0.83	0.82	0.90
F-Measure	0.79	0.79	0.84	0.75	0.73	0.76	0.79	0.79	0.84

Table 3 & Figure 2 represent the ensemble classifiers accuracy before preprocessing. The three ensemble algorithms are employed on data to identify the best algorithm with the highest accuracy. The results shows that SVM with boosting and stacking yields higher accuracy. This greatly helps in the early identification of patients affected with diabetic in an efficient way. The average accuracy for imbalanced class data set is 72.63

Table 4: Accuracy after Preprocessing

Ensemble Classifier (Accuracy %)	Base classifiers		
	DT	KNN	SVM
Bagging	81.60	76.00	78.80
Boosting	84.00	82.00	74.66
Stacking	80.66	82.00	74.66

Table 4& Figure 3 represent the performance of ensemble classifiersaccuracy after preprocessing. The ensemble algorithms are employed on data to identify the best algorithm with the highest accuracy. From the results, it is found that decision tree with boosting provides the highest accuracy on prediction of diabetic disease. The average accuracy for balanced class data set is 79.37

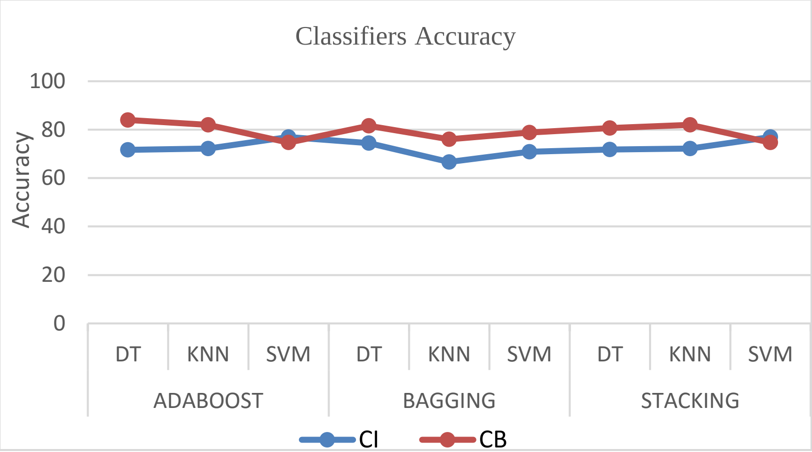

Figure 4: Comparison of Imbalanced classwith Balanced class.

Figure 4 shows the accuracycomparison between imbalanced class and balanced class. From the results obtained balanced class withdecision tree and boostingalgorithm produce the better accuracy.

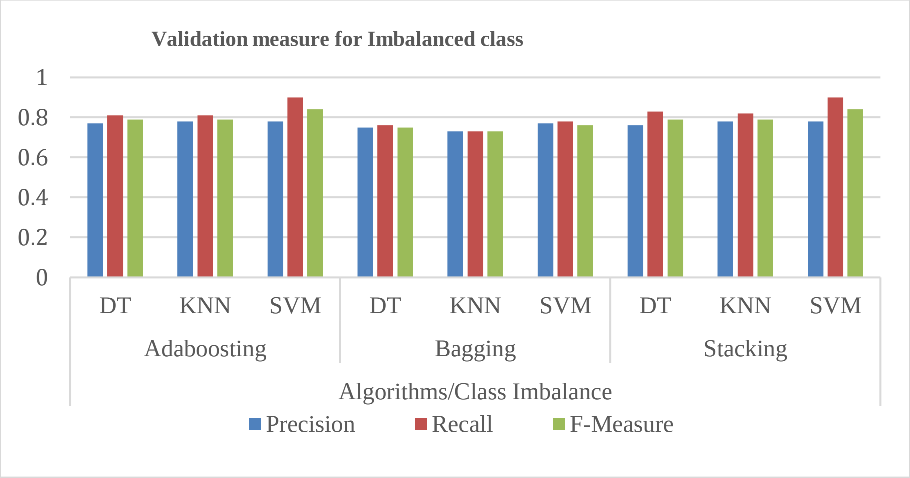

Figure 5: validation measures-before processing

Table 4& Figure 5 represents validation measures of ensemble algorithms. Average precision value for imbalanced class data set is 0.77, the average recall value is 0.81 and average f-measure value is 0.78. Also, ensemble classifiers give better precision values than the single classifiers for imbalanced data [8].

Validation Measure	Algorithms/ Balanced Class								
	Boosting			Bagging			Stacking		
	DT	KNN	SVM	DT	KNN	SVM	DT	KNN	SVM
Precision	0.85	0.80	0.72	0.83	0.82	0.83	0.80	0.80	0.72
Recall	0.79	0.81	0.74	0.83	0.82	0.83	0.77	0.80	0.74
F-Measure	0.82	0.80	0.73	0.83	0.82	0.83	0.78	0.80	0.73

Performance Metrics

The performance metrics can be derived from the confusion matrix. In this work three performance metricsare used namely precision, recall and f-measure.

Precision= TP/ (TP+ FP), Recall= TP/ (TP+ FN)
F-Measure= 2*(Precision*Recall)/ (Precision+ Recall)
Table 4: Validation Metrics

Table 5: Performance Analysis

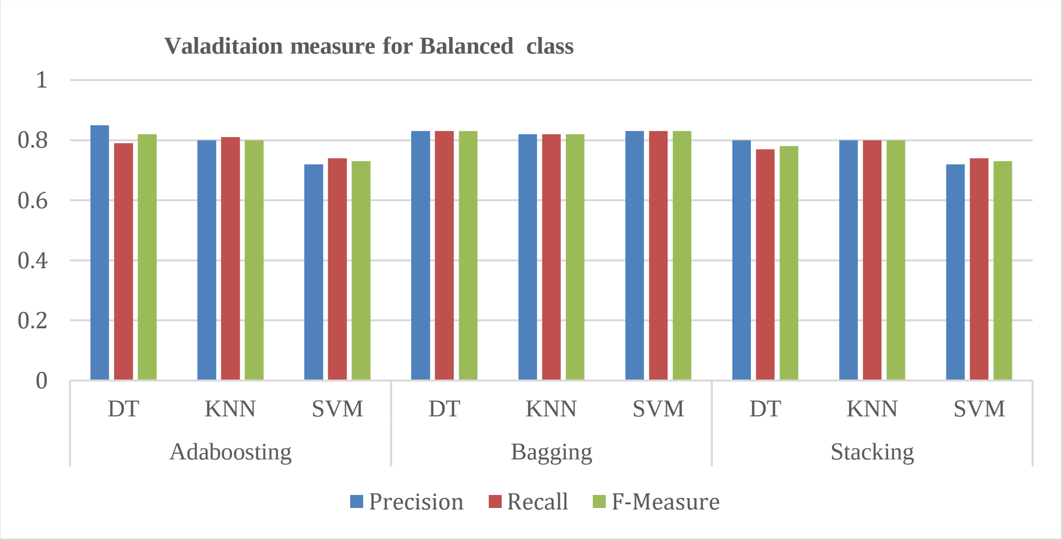

Figure 6: Validation measures-after preprocessing.

Table 5& Figure 6 representsthe validation measures after preprocessing. The average precision value for balanced class data set is 0.80, the average recall value is 0.79 and average f-measure value is 0.79. As the results are clear ensemble algorithm is one of the best suitable algorithms for predicting diabetic disease than the single classifiers.

5. CONCLUSION AND FUTURE WORK

The Proposed model is developed for predicting the presence of diabetic in human. The pre-processing of dataset is being performed bySMOTE. The Boosting, Bagging and Stacking are the classification algorithms which are employed for data classification. Accuracy cum validity measures like precision, recall,f-measure is being utilized for evaluating the performance of machine learning algorithm. By using the above measures we get the result and it’s concluded that decision treewithBoosting combinations ensemble classifier is more suitable for disease prediction.

REFERENCES

[1]. Bhavana, N., Meghana S Chadaga., Pradeep K R. A Review of Ensemble Machine Learning Approach in Prediction of Diabetic Diseases.International Journal on Future Revolution in Computer Science & Communication Engineering.4(3) (2018) 463-466.

[2]. BhondveArti T, BhameVaishali S, KadamAishwarya R, KopnarKomal D, “Breast Cancer Disease Prediction: Using Machine Learning Approach”, International Research Journal of Engineering and Technology, 6(2019).

[3]. Emran Saleh., Jerzy Błaszczynski., Antonio Moreno., Aida Valls., Pedro Romero-Aroca., Sofia de la Riva-Fernandez., Roman Slowinski., Learning ensemble classifiers for diabetic retinopathy assessment. Elsevier - Artificial Intelligence in.Medicine.85 (2018) 50-63.

[4]. Gang Wang, Jinxing Hao, Jian Ma, HongbingJiang,"A comparative assessment of ensemble learning for credit scoring", Elsevier - Expert Systems with Applications, 38(2011) 223-230.

[5]. Herbert F. Jelinek., Jemal H. Abawajy., Andrei V. Kelarev., Morshed U. Chowdhury., Andrew Stranieri. Decision trees and multi-level ensemble classifiers for neurologicaldiagnostics.AIMS Medical Science. 1 (1) (2014) 1-12.

[6]. HimaniBhavsar, Mahesh H. Panchal, "A Review on Support Vector Machine for Data Classification", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), (2012).

[7]. IoannisKavakiotis., OlgaTsave., AthanasiosSalifoglou., NicosMaglaveras., IoannisVlahavas., IoannaChouvarda. Machine Learning and Data Mining Methods in Diabetic Research. Elsevier - Computational and Structural Biotechnology Journal. 15 (2017) 104-116.

[8]. Kalaiyarasi, P., Suguna, J. The Effect of Class Imbalance in Diabetic Disease Prediction by Machine Learning From Healthcare Communities.Journal of Advanced Research in Dynamical & Control Systems.10 (14) (2018) 1135-1141.

[9]. KemalAkyol., Baha Sen. Diabetic Mellitus Data Classification by Cascading of Feature Selection Methods and Ensemble Learning Algorithms.I.J. Modern Education and Computer Science. 6 (2018) 10-16.

[10]. Lidong Wang., Cheryl Ann Alexander. BigData Analytics as Applied to Diabetic Management. European Journal of Clinical and Biomedical Sciences.2(5) (2016) 29-38.

[11]. NiteshV.Chawla, Kevin W. Bowyer, Laurence O. Hall, W. Philip Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intellegence Research 16 (2002) 321-357.

[12]. NongyaoNai-arun., PunneeSittidech. Ensemble Learning Model for Diabetic Classification.Advanced Materials Research. DOI: 10.4028/www.scientific.net/AMR.931-932.1427. 931-932 (2014) 1427-1431.

[13]. PelinYıldırım,Ulaş K. Birant, DeryaBirant, "EBOC: Ensemble-Based Ordinal Classification in Transportation", Journal of Advanced Transportation, (2019).

[14]. PunneeSittidech.,NongyaoNai-arun., Ian T, Nabney. Bagging Model with Cost Sensitive Analysis on Diabetic Data. Information Technology Journal.11 (1) (2015) 82-90.

[15]. Roxana Mirshahvalad., NastaranAsadiZanjani. Diabetic prediction using ensemble perceptron algorithm.IEEE Explore - 9th International Conference on Computational Intelligence and Communication Networks (CICN). DOI: 10.1109/CICN.2017.8319383. (2018).

[16]. Saba Bashir.,UsmanQamar., Farhan Hassan Khan., YounusJaved M. An Efficient Rule-Based Classification of Diabetic Using ID3, C4.5, &amp; CART Ensembles.IEEE Explore - 12th International Conference on Frontiers of Information Technology. DOI: 10.1109/FIT.2014.50. (2015).

[17]. SajidNagi,DhrubaKr.Bhattacharyya. Classification of microarray cancer data using ensemble approach, New model Anal Health Inform Bioinforma, 2:159-173(2013).

[18]. Saravanakumar, N M., Eswari., T. Sampath., P. Lavanya, S. Predictive Methodology for Diabetic Data Analysis in Big Data.2nd International Symposium on Big Data and Cloud Computing (ISBCC’15) - Procedia Computer Science.50 (2015) 203–208.

[19]. Seokho Kang, PilsungKang,TaehoonKo, SungzoonCho,Su-jinRhee,Kyung-Sang Yu, An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction, Expert system with Applications, 4265-4273, Feb 2015.

[20]. SriparnaSaha.,SayantanMitra., Ravi Kant Yadav. A Stack-based Ensemble Framework for Detecting Cancer MicroRNA Biomarkers. Genomics Proteomics Bioinformatics., 381–388, 15(2017).

[21]. Thanga Prasad, S., Sangavi, S., Deepa, A., Sairabanu, F., Ragasudha, R. Diabetic data analysis in big data with predictive method. IEEE Explore - 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET). DOI: 10.1109/ICAMMAET.2017.8186738. (2017).

[22]. Yukai Li., HulingLi.,Hua Yao. Analysis and Study of Diabetic Follow-Up Data Using aData-Mining-Based Approach in New Urban Area of Urumqi,Xinjiang, China, 2016-2017. Hindawi Computational and Mathematical Methods in Medicine. Volume 2018, Article ID 7207151, 1-8.

[23]. Zhiyuan Ma., Ping Wang., ZehuiGao.,Ruobing Wang., KoroushKhalighi. Ensemble of machine learning algorithms using the stacked generalization approach to estimate the warfarin dose.PLoS ONE. https://doi.org/10.1371/journal.pone.0205872. 13(10) (2018) 1-12.

[24]. The Promise of Big Data in Diabetic Management, A thought paper by Scalable Health, March 2017.

[25]. Jiawei Han and MichelineKamber, “Data Mining Concepts and Techniques”, Third Edition, Elsevier, 2012.

[26]. Pang-NingTan,MichaelSteinbach,Vipin Kumar, Introduction to data mining,Pearson Indian Education Service Pvt.Ltd,2016.

[27]. https://www.kaggle.com/uciml/pima-indians-diabetes-database.