ISSN 2347 - 3983
Volume 8. No. 7, July 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter175872020.pdf
https://doi.org/10.30534/ijeter/2020/175872020

Boosting Classifiers in Diabetes Disease Prediction

Shyamili V[1], Shilpa Ankalaki[2]
[1]M Tech Scholar, Nitte Meenakshi Institute of Technology, Bengaluru, India, shyamiliv@gmail.com
[2]Assistant Professor, Nitte Meenakshi Institute of Technology, Bengaluru, India, shilpa.a@nmit.ac.in

ABSTRACT

Diabetes is considered to be one of the most deadly and chronic diseases that causes an increase in blood sugar. Many complications, such as cardiovascular disease, nerve damage, kidney damage, eye damage (retinopathy), etc., may occur if diabetes remains untreated and unidentified. The time-consuming process for the identification of the disease involves the visit of the patient to the diagnostic center, the consultation of the doctor and further clinical testing. But the rise in data mining approaches solves this critical problem of early-stage diabetes prediction. The focus of this paper is to analyze the reliability of different booster classifiers to design a model that can predict the likelihood of diabetes in patients with the highest precision. Therefore, the boosting classifiers AdaBoost, GBM, XGBoost, CatBoost and LightGBM were used in this experiment to detect diabetes at an early stage. Experiments are conducted on the publicly available diabetes database of Kaggle and also the efficiency of the various classifiers is evaluated on the basis of accuracy.

Key words: AdaBoost, GBM, XGBoost, CatBoost, LightGBM

1. INTRODUCTION

More than 246 million people worldwide have been affected by diabetes, with most of them pregnant women. This figure is projected to grow to over 380 million by 2025 according to the WHO report. The disease has been called the fourth deadly illness with no immediate solution in sight in the US. Diabetes cases as well as their symptoms are well documented with the emergence of information technology and its continued advent into the medical and healthcare sector. This paper suggests a quicker and more effective screening procedure, leading to prompt diagnosis of patients. The prevalence of diabetes is rising day by day and is seen more in women than men. Diabetes diagnosis is a stressful operation. And detecting the disease is made possible with progress in science and technology. There is growing emphasis on health care providers to boost their quality and reduce the skyrocketing healthcare costs. Advancing technology and other variables compels healthcare providers to implement innovative communication and coordination structures through their environments[19]. Healthcare providers are now in a position to handle large volumes of digital data in the form of EMR / EHR, medical reports, pharmacy orders, patient reviews and responses. Gestational diabetes during pregnancy is promotes blood sugar level[13]. The placenta-producing insulin-blocking hormones cause this type of diabetes. The data set for diabetes pregnant patients is from the open healthcare dataset in 'Kaggle' is taken both for training and model testing. The final aim is to find solutions for diagnosing the disease by analyzing the patterns found in the data by using the boosting classifiers and analyzing the performance of each classifier. Boosting algorithms is based on those observations of training. There are five commonly used methods of boosting which include AdaBoost, CatBoost, LightGBM, XGBoost, and boosting of gradients. Boosting is a sequential process, where each subsequent model tries to correct the previous model's errors. The successor models depend on the preceding model. The boosting algorithm thus brings together a number of weak learners to form a strong learner.

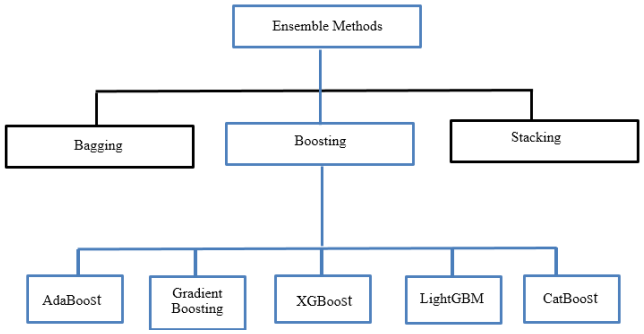


Figure 1: Structural representation of different types of Ensemble Methods

2. RELATED WORK

Sriraam Natarajan et al.[1], According to the study, PPD is not cognitively diagnosable, given the question of determining PPD from demographic, depression, and pregnancy survey data, etc. The paper shows that GB surpasses the most of the existing classifiers. In particular, F3 and F5 scores that offer higher priority to recall, the gradient boosted techniques display statistically considerable improvements in output compared to standard approaches. Yun-peng Wu et al.[2],.This paper uses the AdaBoost algorithm to achieve the weight of technical indicators and to obtain a better prediction result. This work shows that the direct input of stock or future price data to certain ML models does not automatically contribute to successful prediction. The ML model could even achieve high optimization and promotion in this study. Phusanisa Charoen-Ung et al.[3], For predicting an individual plot's sugarcane yield grade two predictive methods have been used here random forest and gradient boosting methods. They examined the methods for recognizing the yield grade by finding the accuracy of the estimation on the test data set which are 71.83% and 71.64% for each. Jinze Li[4], The historical data from monthly rent tags was used to create a LightGBM model based on ML for precise forecasting of monthly housing rent. The experimental error, according to the experimental outcome in the model was with fewer training times and is only 0.1429, and the forecast accuracy is as high as 96 per cent via the LightGBM model. Loshma Gunisetti et al.[5], In this paper, the Decision tree method and the AdaBoost regression method has been employed to apply and create a crop production prediction model. Their inspiration for the examination was to explore possibilities of having open and private mists for groups with light operating burdens. Here the output analysis was done with performance on R-squared score. Sivala Vishnu Murty et al.[6], This paper has tailored an efficient XGBoost model using some of the vital hyper-parameters such as regularization of L2, learning rate, logistic loss function and number of estimators.According to their work it is observed that the model gave an accuracy of 99 per cent which is higher compared to other existing instituted models for classification. Bijun Wang et al.[7], Extraction of appropriate apps and advanced classifiers for machine learning are two types of approaches most often used to identify transportation modes. According this paper, in comparison LightGBM is stronger with XGBoost. The error of the experimental results in the model with fewer training times is just 0.1429 and the prediction accuracy is high. Xiaotong Dou[8], Here a user-data-based behavior assessment system was developed based on user clicks, browses and related product purchases and applied to major online shopping platforms. This study used the CatBoost classifier model applied to the unbalanced data set to find the actual purchase by the buyers. It could achieve 88.51 per cent accuracy and 84.48 recall rate. Anisha.C.D et al.[9], Here Bagging, AdaBoost, GBM and XGBoost are the different classifiers used in this study of PD predictions. Ensemble classifiers are equipped with the optimum parameters provided by the hyper-parameter optimization process, like the random search and grid search. For the proposed approach an accuracy of 94 per cent is attained. Jasmina Novakovic et al.[10], Bagging, AdaBoost, Random Forest and Gradient Boosting Classifier Ensembles are being used for classification purpose. The work shows that the accuracy of fraudulent card purchases correctly identified as fraud transactions, and the accuracy of non-fraudulent card purchases successfully identified as non-fraudulent transactions, is marginally less.

3. PROBLEM STATEMENT

Diabetes is a chronic condition or group of metabolic disorders in which a person suffers from an excessive amount of blood glucose in the bloodstream, which is either deficient in the development of insulin, or because the cells of the body do not respond to insulin. As the name suggests, during pregnancy gestational diabetes appears to develop. Gestational-diabetes is one among other diabetes varieties that impacts how glucose (sugar) is used in the cells of the patient. It induces elevated levels of blood sugar and can affect the pregnancy. This system values AdaBoost, GBM, XGBoost, LightGBM and CatBoost predictive investigation to predict and classify whether or not a patient has diabetes, based on patient's records. This forms two distinct classes, namely diabetic and non-diabetic.

4. DATASET DESCRIPTION

The dataset is taken from publically available Kaggle-diabetes dataset. The data collection is used to identify the symptoms that would affect the woman during their pregnancy. The dataset has the following attributes to it:

i. Plasma Glucose Level- Fasting females with a plasma glucose higher than or equal to 5.1mmol / l (92 mg / dl) but less than 7.0 mmol / l (126 mg / dl) may be diagnosed with gestational diabetes.

ii. Diastolic blood pressure- Regular blood pressure is under 140/90 mm Hg. During pregnancy: Blood pressure between 140/90 and 149/99 mm Hg is slightly elevated.

iii. Triceps Thickness- Around 18 weeks gestation and weight gain in maternal pregnancy between 18 and 28 weeks of pregnancy, maternal triceps skinfold thickness was shallowly inversely related to BP in babies.

iv. Serum Insulin- Resilience to insulin's impact on glucose take-up and use may be associated with stable female pregnancy. IR is defined as an inability of target tissues to respond to normal circulating insulin levels.

v. Body Mass Index (BMI) - BMI calculation for pregnant women will be based on the weight before pregnancy that can stated as: Good weight = 18.5 to 24.9, Overweight = 25 to 29.9 and 30 to 39.9 = obese.

vi. Diabetes Pedigree- High levels of glucose in blood during pregnancy may also increase the likelihood of your baby being

born early, weighing more, or having breathing problems or low glucose in blood right after birth.

vii. Age-The risk of GDM will increase significantly and gradually from 25 years on. Using age as the cutoff for screening for about 25 years, and finding that the most predictive component of GDM has been maternal age for about 25 years.

viii. Pregnancies- The pregnancy count also may increase the risk level of the patient in getting diabetes, especially if the patient had a history of having gestational diabetes in the previous pregnancy period.

5. PROPOSED SYSTEM

This paper aims at comparing the results of the AdaBoost, GBM, XGBoost, LightGBM and CatBoost classifiers in diabetes prediction. Here the various parameters obtained are grouped into descriptive groups to show their meaning with regard to the Diabetes or Non Diabetes labels. This will facilitate us to signify the group of parameters be appropriate to either of the two classes. This will also have the parameters statistical significance. Figure 2 shows the flow of the proposed work.

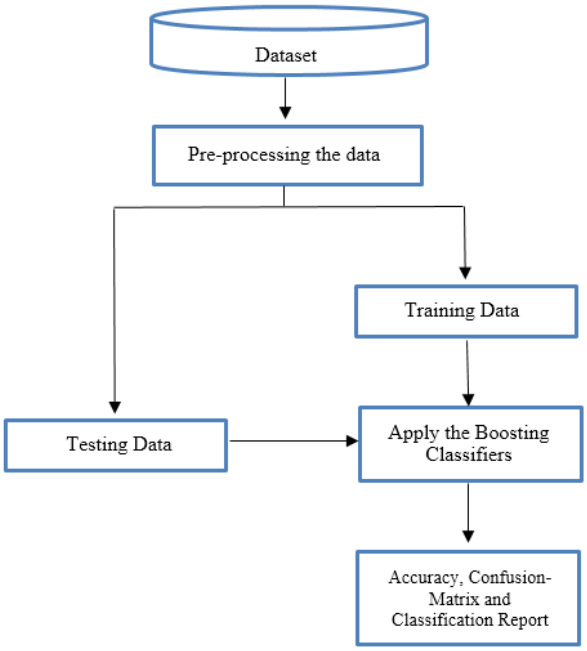


Figure 2: Overview of proposed approach

Step 1: Preprocessing the input dataset

As a first step in cleaning the data, the dataset was checked for any missing or null values. There were no presence of null or empty values in the dataset. Next task here was to find if there were any outliers in the data. For this purpose a join grid plot was plotted for each feature in the dataset. From the grid plot it was observed that the features Blood-Pressure, BMI, Glucose and Skin-Thickness were having values zero, which is incorrect. Removal of outliers from the dataset was done by directly eliminating the outlier values from the dataset and thus a modified data was further given as input which was then split into training and testing data.

Step 2: Use of grid search in tuning the hyper-parameters

Tuning the hyper-parameters plays a very important role in creating the model as it can make or break the models. To obtain the best set of values for the parameters GridSearchCV has been used. Grid search will train a model with all the possibilities of combinations of the hyper-parameters and gives the best out of it. Based on the results from the grid search the parameter and the values that it was tuned to are listed in Table 1.

Step 3: Fitting the classifiers

Diabetes prediction analysis is based on the classifier AdaBoost, GBM, XGBoost, CatBoost, and LightGBM which have been used with optimal values calculated through the method of tuning hyper-parameters.

A. AdaBoost

One of the simplest boosting algorithms is the AdaBoost (Adaptive Boosting) classifier. Initially, AdaBoost assigns equal weights to each observation of training. This uses several weak models and assigns higher weights to certain results that have been observed for misclassification. Combining the effects of the decision boundaries reached through multiple iterations, because it uses multiple weak models. The accuracy of the misclassified observations is improved, thereby also improving the accuracy of the overall iterations.

Pseudocode for AdaBoost Algorithm:

Let the initial weights be W: W_1, W_2,..,W_n=(1/n) and C be the weak classifier.

For each value of i in [1,C],

Align the weak classifier WC^i with each sample weight W

Error, $E^i = ((\sum^n_{j=1} W_j \, l(WC^i(X_j) \neq Y_j)) / (W_j)$ (1)

For coefficient of WC^i,

$$\alpha^i = \log((1- E^i) / E^i) + \log (K - l) \quad (2)$$

If wrong, then increase the weight,

for every W_j in W:

$$W_j = W_j * e^{\wedge}(\alpha^i * l(WC^i(X_j) \neq Y_j)) \quad (3)$$

To normalize the weight, W = W – mean (W) (4)

The final output of the model is taken by weighted voting (i.e to find the class with the highest vote)

Prediction, $Y_j = \max_k(\sum^i_{i=1} \alpha^i \, l(WC^i(X_j) = k))$ (5)

The AdaBoost classifier trains a series of models with increased sample weights, producing Alpha confidence

coefficients for individual classification models on the basis of error. Low errors result in a large Alpha, which indicates greater importance in the voting process.

B. GBM

In 2001 Friedman introduced GBM, short for "Gradient Boosting Machine". GBM uses the boosting technique to create a strong learner, integrating a number of weak learners[11]. Regression trees used as a base learner, each subsequent tree in series is based on the previous tree's measured errors. It creates new models sequentially from an ensemble of weak models, with the idea that each new model can minimize the function of loss.

The pseudocode of the basic working of GBM classifier:

To improve the model 'F', We have to minimize the loss, i.e minimize L(Y,F(X))

The loss function, L = func(F(X_1), F(X_2),…, F(X_n),Y)

The minimization is performed by integrating an estimator, E on (X_i , $\frac{\partial L}{\partial X^i}$) $\forall i$

F(X) + E(X) is an approximation of the gradient-descent,

$$F(X_i) = F(X_i) - \frac{\partial L}{\partial F(Xi)} \quad (1)$$

Let WC be the weak classifier.

To integrate the estimator F^1,

for each i in [1,WC],

$$\text{Loss, } L^i = \sum_{j=1}^{n} (Y_j - F^i(X_j))^2 \quad (2)$$

To calculate the negation gradient,

$$-\frac{\partial Li}{\partial Xj} = -\left(\frac{2}{n}\right) * (Y_j - F^i(X_j)) \forall i \quad (3)$$

Integrate the weak estimator, E^i on (X , $\frac{\partial L}{\partial X}$)

Let $'\rho'$ be the change in step size, then Prediction,

$F^m(X) = F^i(X) + \rho * E^i(X) = F^1 + \rho * \sum_{i=1}^{m} E^i(X)$

C. XGBoost

XGBoost is an algorithm based on a decision-tree, using a gradient-boosting framework. Both the ensemble tree methods, XGBoost and GBM apply the principle of strengthening the poor learners[16]. XGBoost divides in accordance to the specified hyper-parameter of max-depth, and after that begin the reverse pruning of the tree and delete breaks below that and no positive gain is obtained.[14]. This strategy is used since a split without minimizing loss can often be followed by a split of a minimizing loss.

The XGBoost algorithm:

Make initializations for the value of $f_o(x)$

For all r=1, 2,…, P do

$$\text{Compute } g_r = -\frac{\partial L(y,F)}{\partial f} \quad (1)$$

$$\text{Compute } h_r = -\frac{\partial^2 L(y,F)}{\partial f^2} \quad (2)$$

Compute the structure (by choosing the splits which has max-gain),

$$M = \frac{1}{2}[(G^2_L / H_L) + (G^2_R / H_R) - (G/H)] \quad (3)$$

Calculate the leaf-weights,

$$w^* = -\frac{G}{H} \quad (4)$$

Calculate the base learner,

$$\hat{a}(x) = \sum_{j=1}^{T} wI; \quad (5)$$

Add the trees,

$$f_r(x) = f_{r-1}(x) + \hat{a}(x); \quad (6)$$

end

The Result, $f(x) = \sum_{r=0}^{P} f_r(x)$ (7)

D. CatBoost

CatBoost is a gradient boosting implementation, using binary decision trees as reference predictors. CatBoost offers the new Minimal Variance Sampling (MVS) methodology. In this technique, weighted sampling occurs at the tree level and not at the split level. A balanced tree generated using CatBoost. In each level of such a tree, the feature-split pair that leads to the lowest loss is chosen, and is used for all level nodes[15].

Building a tree using CatBoost-algorithm:

Let the input be C,$\{y_i\}^n_{j=1}$,a, L, $\{\sigma_i\}^t_{j=1}$, Mode

grd = Calculate_Gradient(L,C,y);

r = random_func(1,t);

if Plain Mode then

$$G = (grd_Y(1),\ldots, grad_Y(n))$$

if Ordered Mode then

$$G = (grd_{Y,\sigma Y(1)-1}(j) \text{for } j=1 \text{ to } n)$$

T = empty tree;

For each step of down procedure do

For each candidate split c do

T_c = add the split c to T;

if Plain Mode then

$\Delta(j)$ = average(grd_Y(q) for q: leaf(q) = leaf(j)) $\forall j$

if Ordered Mode then

$\Delta(j)$ = average($grd_{Y,\sigma Y\,(i)-1}$(q) for q: leaf(q) = leaf(j), $\sigma_Y(q) < \sigma_Y(i)$) $\forall j$

end

$$loss(T_c) = ||\Delta - G||_2$$

end

$T = arg\_min_{Tc}(loss\ (T_c))$

if Plain Mode then

$C_{r`}(j) = C_{r`}(j)$ a average($grd_{r`}$(q) for q: leaf(p) = leaf(j)) $\forall r`, j$)

if Ordered Mode then

$C_{r`,j}(j) = C_{r`,k}(j)$ a average($grd_{r`,k}$(q) for q: leaf(q) = leaf(j), $\sigma_{r`}(q) \leq k$) $\forall r`, k, j$;

return (T,C)

Here the plain mode would be a combination of the regular GBDT method and an orderly target metric. In ordered mode it implements a randomized permutation of the training samples - $\sigma 2$, and Hold n distinct supportive models - $C1, \ldots, Cn$ such that model Ci is trained on possible combination using just the first i samples.

E. LightGBM

LightGBM is a GBDT-based data model which Microsoft introduced in 2017[12]. The algorithm uses a leaf-wise generation strategy to reduce training data[18]. LightGBM offers one-sided gradient sampling (GOSS) that extracts a split using all high gradient (i.e. large error) instances and a randomly chosen set of smaller gradient errors. By raising the amount of data instances and maintaining the precision of qualified decision trees, GOSS strikes a fair balance with the size.

The GOSS algorithm can be showed as:

Let D be the training-data, n be the number of iterations, x be the sampling ratio of the huge gradient data, y be the small gradient-data sampling ratio, loss be the loss-function, and L be the weak learner.

Let models = {} and val = $\frac{1-x}{y}$

```
Top_N = x × len(D)
Rand_N = y × len(D)
for each i from i = 1 to n do
    pred = models.predict(D)
    g = loss(D, pred) and w = {1,1,...}
    sorted =GetSortedIndices(abs(g))
    Top_Set =sorted[1:Top_N]
    Rand_Set = RandomPick(sorted[Top_N:len(D)] Rand_N)
    Used_Set = Top_Set + Rand_Set
    val = val × w[Rand_Set]
```

Allot the new weight val to the smaller gradient data.

```
New_Model = L(D[used_Set], − g[used_Set], w[used_Set])
models.append(New_Model)
```

Initially, the algorithm sorts the input data instances as per their absolute gradient value and then picks the data-instances of the top x×100 per cent. Next, arbitrarily samples the y×100 percent data-instances from rest of the input data. The algorithm then magnifies the sampled data with tiny gradients by a constant $\left(\frac{1-x}{y}\right)$, while calculating the information gain.

In doing so, it provides better attention to the under-trained cases without much altering the original distribution of data.

Step 4: Results prediction

The test outcomes are estimated as 0 and 1 where Non-Diabetic is defined by 0 and Diabetic is defined by 1.

Step 5: Evaluation of prediction

The model is evaluated using performance metrics such as accuracy, confusion matrix, and classification report. Confusion Matrix is used to portray the values of true-positive, true-negative, false-positive and false-negative. Accuracy is the percentage of circumstances that have the correct description to counting total circumstances. In the classification report the visualisation displays the model's accuracy, recall, F1, and support scores.

Step 6: Apply K-Fold Validation

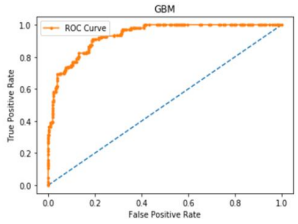
This functions by dividing the dataset into k-parts so that each break of data is called a fold. The classifier must be trained with one kept down on k-1 folds and checked upon on retained fold. It has to be replicated such that each fold of data has an opportunity to be carried back to the testing set. Here the value of K used is 5. The outcome is a more accurate estimate of the performance of the new data algorithm based on the test results. This is more reliable because the algorithm is trained and tested several times on a variety of data.

6. EXPERIMENTAL RESULTS

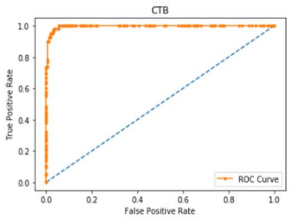
Comparison of performance is enlisted based on the accuracy of the classifiers used on the predicted diabetic data. A confusion matrix (refer figure 3), classification report and ROC plot (refer figure 2) of the classifiers has also been formulated in prediction. From the Accuracy graph plot in figure 4 and 5 it is evident that LightGBM provides highest accuracy in prediction compared to the rest of the classifiers. The confusion matrix and the classification report that were generated also supports the observation that for the current dataset LightGBM outperforms the rest of the classifiers in diabetes prediction.

Table 1: Parameter tuning and difference in accuracy because of the same

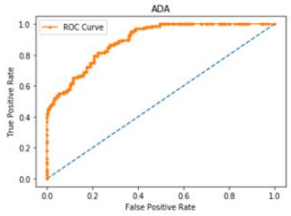
	Ada-Boost	GBM	XGBoost	LightGBM	CatBoost
Accuracy before tuning	0.81	0.833	0.821	0.97	0.96
Parameters Tuned	n_estimators =200	n_estimators = 400 learning_rate = 0.1 min_samples_split = 500 min_samples_leaf = 50 max_depth = 8 max_features = sqrt subsample = 0.8	n_estimators = 200 learning_rate = 0.1 max_depth = 4 min_child_weight = 6 colsample_bytree = 0.8 reg_alpha = 0.005 nthread = 4 subsample = 0.8	n_estimators = 200 learning_rate = 0.15 max_depth=25 min_child_samples = 111 min_child_weight = 0.01 num_leaves = 38 reg_lambda = 0.1 subsample = 0.303	learning_rate = 0.15 depth = 10 eval_metric = AUC iterations = 500 l2_leaf_reg = 9
Accuracy after tuning	0.8347	0.8665	0.8453	0.9809	0.9767
Accuracy on K-Fold Validation	0.841	0.8798	0.8586	0.9931	0.9873



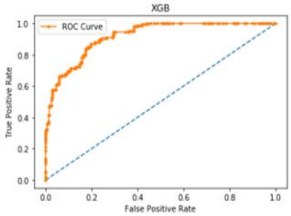
(a) GBM



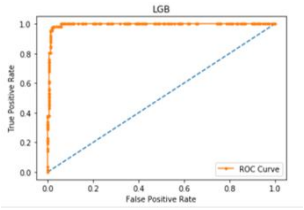
(b) CatBoost



(c) AdaBoost

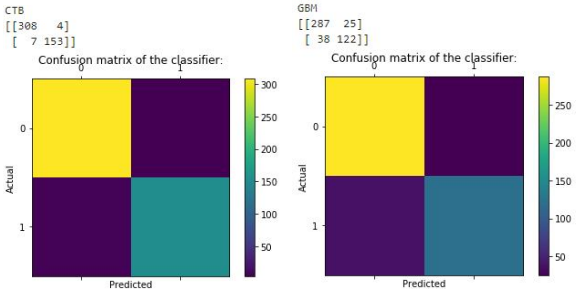


(d) XGBoost



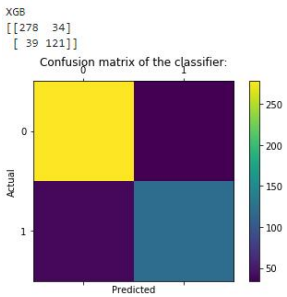
(e) LightGBM

Figure 3: ROC for the respective classifiers

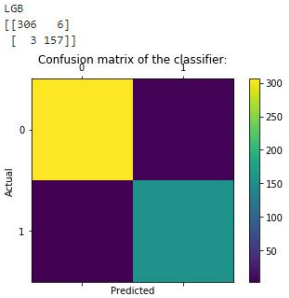


(a) CatBoost

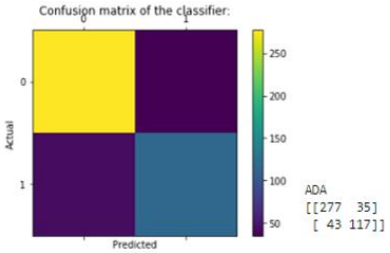
(b) GBM



(c) XGBoost

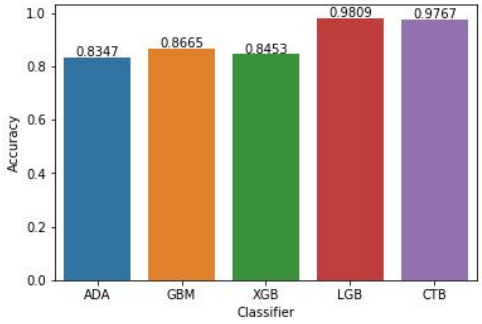


(b) LightGBM

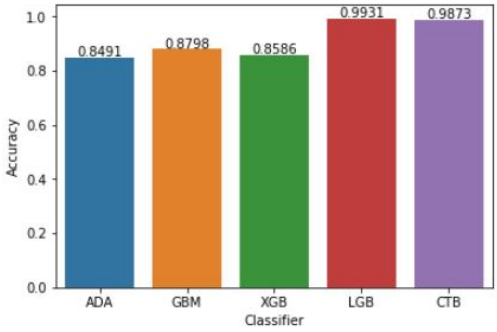


(e) AdaBoost

Figure 5: Confusion Matrix generated for the respective classifiers



(a)



(b)

Figure 4: Accuracy bar-graph obtained for comparison on respective classifiers (a) Before applying K-Fold and (b) After applying K-Fold

7. CONCLUSION

This is a comparative study of the classifiers AdaBoost, XGBoost, LightGBM, CatBoost and GBM in diabetes prediction. The classification was performed based on the number of attributes or features present in the dataset of the sample input. Effectively, the results obtained were represented using graphs. Since the same input data was provided for AdaBoost, XGBoost, LightGBM, CatBoost and GBM classifiers, it was easy to compare the results.

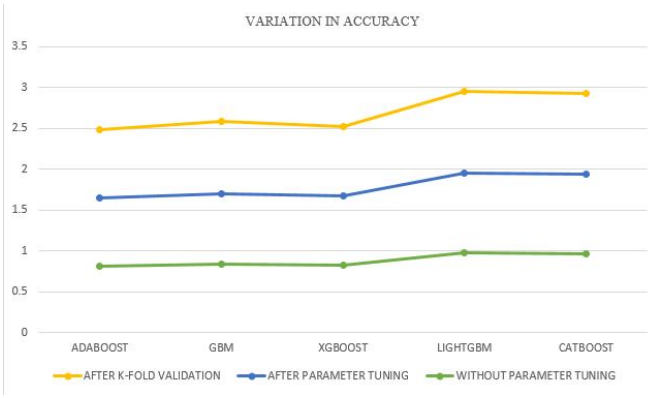


Figure 6: Trend observed in variation of accuracy in respective classifiers

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to Prof. N.R Shetty, Advisor and Dr. H.C Nagaraj, Principal, Nitte Meenakshi Institute of Technology for providing constant support in carrying out the research.

REFERENCES

1. Sriraam Natarajan, Annu Prabhakar, Nandini Ramanan, Anna Baglione, Kay Connelly, Katie Siek, **Boosting for Postpartum Depression Prediction**, *IEEE*, 2017. https://doi.org/10.1109/CHASE.2017.82
2. Yun-peng Wu, Jin-min Mao, Wei-feng Li, **Predication of Futures Market by Using Boosting Algorithm**, *IEEE*, March 2018.
3. Phusanisa Charoen-Ung and Pradit Mittrapiyanuruk, **Sugarcane Yield Grade Prediction using Random Forest and Gradient Boosting Tree Techniques**, *IEEE*, July 2018.
4. Jinze Li, **Monthly Housing Rent Forecast based on LightGBM (Light Gradient Boosting) Model**, *International Journal of Intelligent Information and Management Science* (2018).
5. Loshma Gunisetti, Shirin Bhanu Koduri, Ch Raja Ramesh, K V Mutyalu[1] and D. Ganesh, **Prediction of crop production using AdaBoost regression method**, *Journal of Physics: Conference Series s 1228* (2019). https://doi.org/10.1088/1742-6596/1228/1/012005
6. Sivala Vishnu Murty and R Kiran Kumar, **Accurate Liver Disease Prediction with Extreme Gradient Boosting**, *IJEAT*, August 2019.
7. Bijun Wang et al., **Detecting Transportation Modes Based on LightGBM Classifier from GPS Trajectory Data**, *IEEE*, 13-February-2020.
8. Xiaotong Dou, **Online Purchase Behaviour Prediction and Analysis Using Ensemble Learning**, *IEEE 5th International Conference on Cloud Computing and Big Data Analytics*, 2020.
9. Anisha.C.D and Dr. Arulanand.N, **Early Prediction of Parkinson's Disease (PD) Using Ensemble Classifiers**, *International Conference on Innovative Trends in Information Technology* (ICITIIT), 2020.
10. V. Mareeswari**, Prediction of Diabetes Using Data Mining Techniques**, *Research Journal of Pharmacy and Technology 10(4):1098*, January 2017. https://doi.org/10.5958/0974-360X.2017.00199.8
11. https://towardsdatascience.com/understanding-gradient-boosting-machines.
12. Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu, **LightGBM: A Highly Efficient Gradient Boosting Decision Tree**, *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, December 2017.
13. https://www.webmd.com/diabetes/gestational-diabetes.
14. Tianqi Chen and Carlos Guestrin, **XGBoost: A Scalable Tree Boosting System**, *Association for Computing Machinery, KDD '16*, August, 2016.
15. Liudmila Prokhorenkova et al., **CatBoost: unbiased boosting with categorical features**, *32nd Conference on Neural Information Processing Systems (NeurIPS)*, 2018.
16. Deok-Kee Choi, **Data-Driven Materials Modeling with XGBoost Algorithm and Statistical Inference Analysis for Prediction of Fatigue Strength of Steels**, *International Journal of Precision Engineering and Manufacturing*, 2019.
17. Pedro Carmona, Fransisco Climent and Alexandre Momparler, **Predicting bank failure in the U.S. banking sector: An extreme gradient boosting-approach**, *International Review of Economics & Finance 61:304-323*, May 2019.
18. Madam Chakradar and Alok Aggarwal, **A Machine Learning Based Approach for the Identification of Insulin Resistance with Non-Invasive Parameters using Homa-IR**, *International Journal of Emerging Trends in Engineering Research(IJETER)*, Vol. 8, 5 May 2020. https://doi.org/10.30534/ijeter/2020/95852020
19. Abhale Babasaheb Annasaheb and Vijay Kumar Verma, **Data Mining Classification Techniques: A Recent Survey**, *International Journal of Emerging Technologies in Engineering Research (IJETER)*, Volume 4, Issue 8, August (2016).