IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Implementation Of Hybrid Ensemble Gradient Boosting Algorithm To Predict Diabetes Health Care Analytics

[1]Deepa. S, [2]Dr. B. Booba

[1]Research Scholar (UP19P9611053), [2]Professor

[1]Department of Computer Science,

[1]Vels Institute of Science, Technology & Advanced Studies(VISTAS), Pallavaram, Chennai-600117, India

***Abstract*:** Diabetes is a chronic disease where the blood sugar levels in human body elevates and leads to serious damages to many organs such as heart, eyes, nerves system and kidney of our body. Healthcare Analytics uses several data analysis methods and techniques to improve patients care and healthcare administration. For this reason, several machine learning algorithms were used to build a machine learning model to predict the disease at the earliest to prevent it. The research work is carried out for 770 diabetes patients of Andaman & Nicobar Islands. The patients are registered and diagnosed in Diabetic Care Clinic, Port Blair, Andaman & Nicobar Islands. The research work uses exploratory data analysis methods for data preprocessing and splits the dataset into training and testing set. Then the research work used feature engineering techniques to identify the importance of each and every attributes and predict the diabetes mellitus accurately based on IDRS (Indian Diabetes Risk Score) risk factor. The existing model of this research work uses nine machine learning algorithms such as XG Boosting classifier, Ada Boosting Classifier, Bagging Classifier, Random Forest Classifier, Logistic Regression, Linear SVC Algorithm, Gaussian Naïve Bayes Algorithms, KNN classifier and decision trees algorithm and produce accuracy, precision, recall and f1 score. From the results, it is came to know that Bagging classifier, Ada Boosting, XG Boosting classifier, KNN Classifier and Random Forest has produced the highest accuracy as 88% followed by Decision Trees, Logistic regression and Gaussian Naïve Bayes algorithm. Based upon these results, the research work selects XG Boosting, Ada Boosting and Bagging Boosting Classifier and combines it using ensemble technique and develop hybrid ensemble gradient boosting algorithm and implemented it in the proposed system and produced the results. From these results, it is learned that the proposed ensemble gradient boosting algorithm outperforms rest of the machine learning algorithms and produced best accuracy of 99%, precision of 79%, recall of 84%, Area Under Curve of 87% and ROC as 87%. Based on this, it is concluded that the proposed hybrid ensemble gradient boosting classifier not only outperforms rest of all the machine learning algorithms and produce better results but also accurately predicts diabetes mellitus of the patients based on risk factors.

Keywords: Accuracy, AUC, Cross Validation Score, Diabetes Mellitus, F1 Score, Hybrid Gradient Boosting Algorithm, Precision, Recall and ROC Curve

1 Introduction

Diabetes Mellitus is a group of metabolic disease in which there are high blood sugar levels for a prolonged period in human body. If diabetes left untreated, it may leads to serious complications or even in mortality. Because of this, several machine learning algorithms were used to build a predictive model to detect and predict the disease at the earliest to prevent it.

The research work uses supervised machine learning techniques which shows how Hybrid ensemble Gradient Boosting algorithm as compared to various machine learning algorithms like Linear SVC Algorithm, Gaussian Naïve Bayes algorithm, Logistic regression Classifier algorithm, KNN Classifier, XG Boost Classifier, Light GBM and Random Forest Classifier algorithm plays a vital role for diabetes prediction along with the parameters of Accuracy, Precision, f1 score, recall, cross validation, AUC and ROC curve. The research work uses ensemble techniques and confusion matrix to produce the better results.

The research work is carried out for 770 diabetes patients of Andaman and Nicobar Islands and the data has been collected from Diabetes Care Clinic, Junlighat, Port Blair, Andaman & Nicobar Islands.

2. Scope of the Research Work

Diabetes is a most dangerous disease which may cause severe damages to kidneys, eyes, heart and nervous system if it's left untreated. It is learned that the use of emerging IT systems in medical sciences to diagnose, prevent and eradicate life-threatening diseases such as Diabetes, cancer, etc. garnered the attention of IT researchers worldwide. So, several researchers have used many machine learning algorithms to detect and predict the diabetes mellitus. The main aim of the research work is to develop hybrid machine learning algorithm for predicting diabetes mellitus based on certain risk factors and check whether the particular patient is suffered from diabetes or not and also provide better accuracy and precision by reducing the error rates as compared to rest of the work done so earlier. Therefore, this research will have more scope to carry out the research work.

3. Objectives of the Research Work

With the help of hybrid ensemble gradient boosting algorithm along with confusion matrix, the main aim of this research is to predict diabetes healthcare data analytics and provide better accuracy score, precision score, recall score, f1 score, AUC and ROC curve score along with better cross validation score as compared to rest of the machine learning algorithms. Based on these, the aims and objectives of the research work are stated in terms of Primary Objectives and Secondary Objectives.

- The primary objective of the research is to predict diabetes mellitus of Andaman & Nicobar Islands patients based on certain risk factors using proposed hybrid machine learning model.
- The secondary objective of the research work is to provide best Accuracy, Precisions, Recall, F1 Scores, AUC Score, ROC Curve and Cross Validation Score using hybrid ensemble gradient boosting algorithm.

4. Literature Review of the Research Work

In this research work, several papers have been collected related to supervised machine learning algorithms for detecting and predicting diabetes mellitus. Thus, this work provides the various kinds of machine learning algorithms used for early detection and predicting the risk factor of various kinds of diabetes and produce better accuracy and precision score as compared to rest.

Mohammad Ehsan et al.[1] uses support vector machine (SVM), random forest (RF), multi-layer perceptron (MLP), and k-nearest neighbor (KNN). The outcomes shows the SVM model's best performance, accuracy of 98.78%, specificity of 99.28%, and sensitivity values of 97.32% across 50 iterations.

Ashisha GR et.al. [2] has used Light Gradient Boosting Machine (LightGBM), Gradient Boost classifier (GBC), and Random Forest (RF) algorithm in their research work and the findings showed that the combination of the ensemble Voting Classifier and Boruta feature selection algorithm performed better with an accuracy of 90% and 93% respectively.

Shahid Mohammad Ganie et. al. [6] has conducted experiments on five boosting algorithms on the Pima diabetes dataset. Gradient boosting achieved the highest accuracy score of 92.85% among all other algorithms.

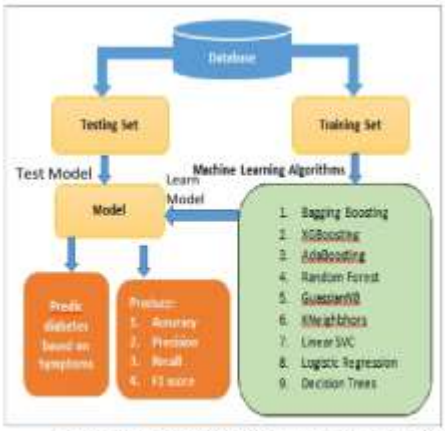
Michael Onyema Edeh et.al.[10] uses four machine learning classification algorithms, namely supervised learning algorithms (Random forest, SVM and Naïve Bayes, Decision Tree DT) and unsupervised learning algorithm (k-means) to identify diabetes in its early stages.

Sasmita Padhy et.al. [11] propose a noninvasive self-care system based on the IoT and machine learning (ML) that analyses blood sugar and other key indicators to predict diabetes early. The main aim of their study is to develop enhanced diabetes management applications which help in technology-assisted decision-making and patient monitoring. By combining both the bagging and boosting methods, the proposed hybrid ensemble Machine Learning model predicts diabetes mellitus.

Rishab Bothra [12] proposed a diabetes prediction model using Machine Learning algorithm for better classification prediction. He used different Machine Learning algorithms to find which gives the better accuracy of classification.

5. Methodology of the Research

The research work uses ensemble technique to build a machine learning model and produce better results using gradient boosting algorithms. For this reason, 770 datasets have been collected from Diabetes Care Clinic, Junglighat, Port Blair, A & N Islands. After collecting the datasets, it has been loaded in the machine learning model and then cleaned and splits into training and testing test. There are 229 datasets have been used for training whereas 553 datasets have been used for testing the data. For training and testing the datasets, ensemble gradient boosting classifier, Light GBM classifier and Random Forest Classifier machine learning algorithms has been used. The architecture of the existing and proposed system is shown in following figures.



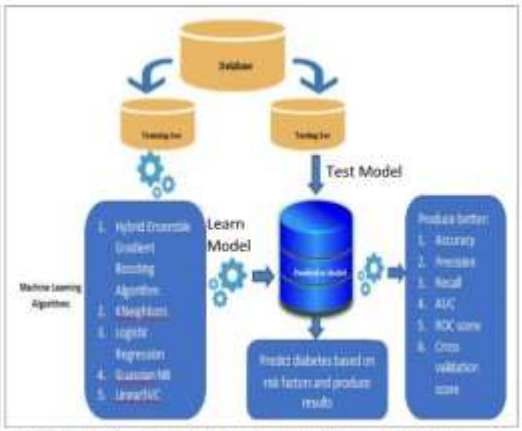


Figure 1: Existing System Model **Figure2:** Architecture of Proposed System Model

The research work has used classification model along with confusion matrix to predict diabetes mellitus and produce better accuracy score, precision, f1score and recall. The research work uses following calculation to find out the best accuracy score, precision, f1 score and recall.

$$\boldsymbol{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}$$

$$\boldsymbol{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$\boldsymbol{Error\ Rate} = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}$$

Where,

TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

5.1 Tools Used

This research has used colab.research.google.com along with Pandas for implementing all the machine learning algorithms to predict the diabetes mellitus and produce accuracy, precision by reducing error rates.

Both Colab and Pandas can take data from a wide range of sources such as .csv files, text files, etc. which is stored in Google drive.

5.2 Exploratory Data Analysis

This is the first phase of the model where the collected datasets are loaded into the database and cleaned and perform several operations using numpy, mathplotlib, seaborn, plotly, etc. For Example, The research work gives the number of diabetic and non-diabetic patient.

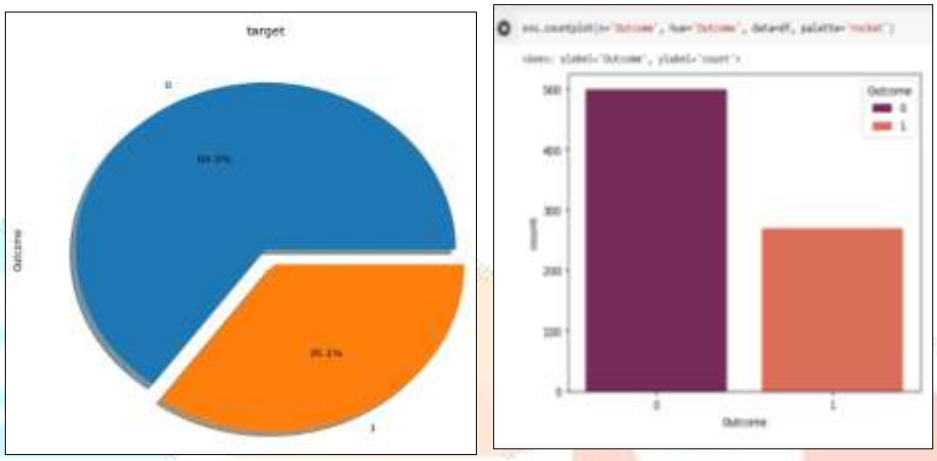


Figure 3: Diabetic vs Non-Diabetic Patients

In the above figure 3, the value ‘0’ indicates non-diabetic patients and the value ‘1’ indicates diabetic patients. From the above figure, it is also learned that there are 35% patients (270) are suffered from diabetes mellitus whereas 65% patients (500) are non-diabetes.

Apart from this, this research also gives the missing dataset details in matrix form. These are shown in following figure.

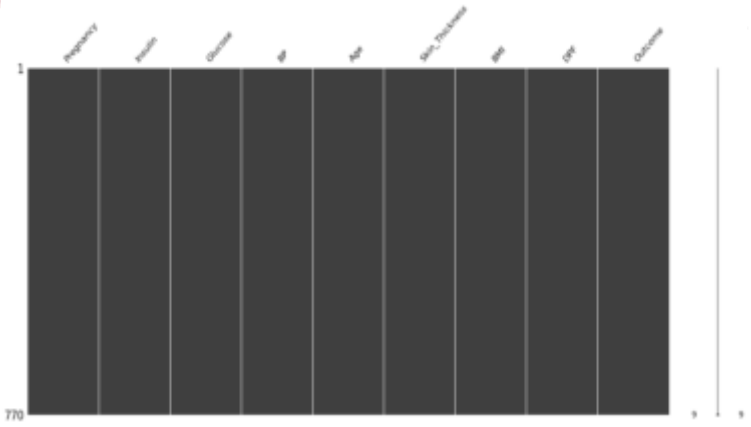


Figure 4: details of Missing Data

From the above figure 4, it is clearly visible that there are no missing values in the collected dataset. The research work will also show the correlation between each and every instances. These are depicted in following figure.

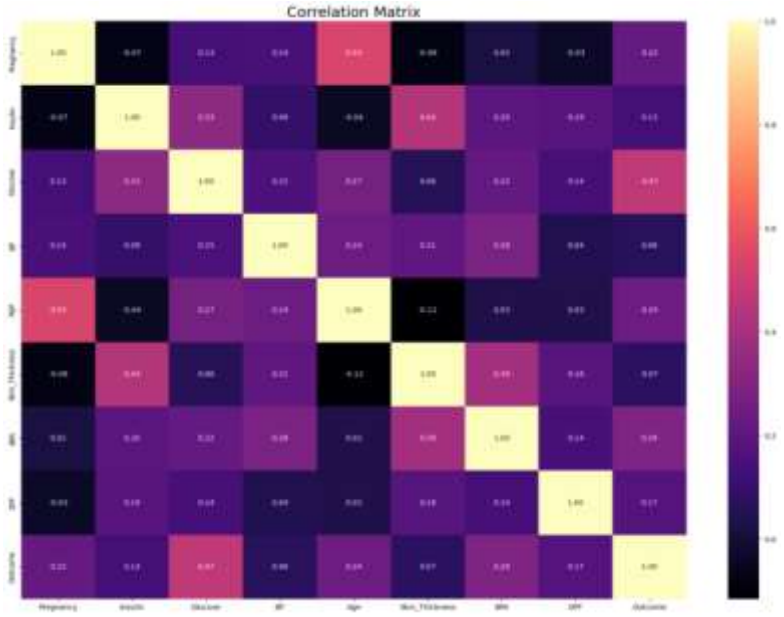


Figure 5: correlation of datasets

From the above figure 5, it is clearly visible that each and every instance has correlation with each other.

6 Building Predictive Model

Step 1:Firstly, the resaerch has defined a problem.

Step 2: Then the proper dataset has been fetched. In the research work, real time datasets are used. The data are collected from Diabetic Care Clinic, Port Blair, Andaman and Nicobar Islands.

Step 3: After collecting the data, the research work has applied the EDP (Exploratory Data Analysis) for graphical representation and also to find out the correlation among each and every instances.

Step 4: Then the research work has applied Feature Engineering techniques to find out the features of each and every instances.

Step 5: After that the research work has splits the dataset into Training and Testing Dataset.

Step 6: After splitting the dataset into training and testing set, the Predictive Machine Learning Model has been built and saved for later use.

Step 7: After building a model, all the machine learning algorithms along with hybrid gradient boosting machine learning algorithm has been implemented using SCIKIT learn and ensemble technique.

Step 8: After implementing machine learning algorithms, Performance Evaluation has been done for the parameters - Accuracy, F1 Score, Precision, Recall, Cross Validation Score, AUC score and ROC curve using confusion matrix.

7 Data Preprocessing and Feature Engineering Techniques

To find the importance of each and every instances, feature engineering techniques have been used. After finding the importance of each and every instances, the datasets are splits into training and testing set. There are 533 training and 229 testing datasets along with 8 attributes and 1 target variable. These are shown in following figure:

```
[ ]  data_train.shape
     (533, 8)

[ ]  target_train.shape
     (533,)

[ ]  data_test.shape
     (229, 8)

[ ]  target_test.shape
     (229,)
```

Figure 6: Training and Testing Data

After splitting the dataset, the research work has used feature importance techniques. It is a technique which is used to calculate the scores for all the input instance or attributes of a predictive system model. Here, the scores represent the importance of each attribute or feature. Higher the score means, the specific feature will have a larger effect on the predictive model that is being used to predict a certain feature or variable. For getting the feature importance of a diabetes dataset, the research work has used ensemble random forest classifier's feature importance technique. The following result has been produced after applying the above technique.

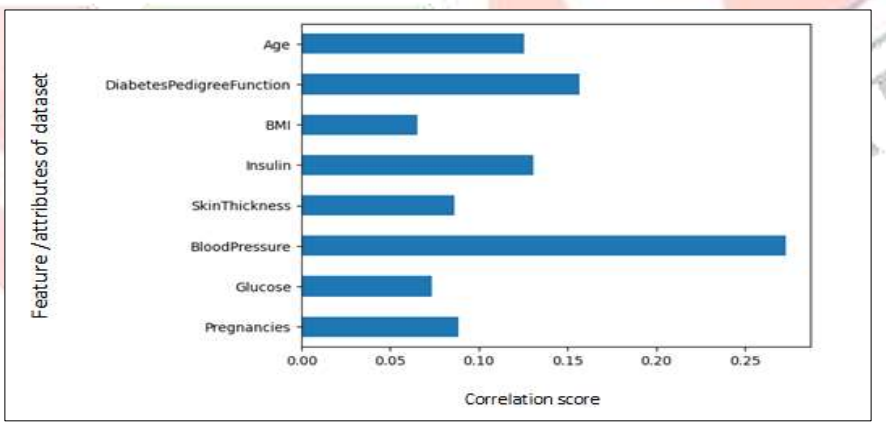


Figure 7: Feature Importance of a dataset

From the above figure 7, we can clearly see that blood pressure as a feature has the highest importance in this dataset followed by Diabetes Pedigree Function, insulin level and age of the patient.

After implementing feature importance techniques, the research work has saved the predictive model for later use. To save a machine learning model, the research work has imported pickle. After that the research work has loaded the saved model to make predictions. To predict whether the patients has diabetes mellitus or not the model has used model.predict(values) method. The following figures shows how the model has been saved, loaded and predicted the diabetes mellitus.

```
[ ] rfc.predict([[6,171,148,72,50,35,33.6,0.627]]) #1st patient
    array([1])

[ ] rfc.predict([[0,168,137,40,33,35,43.1,1.19775]]) #4th patient
    array([1])

[ ] rfc.predict([[5.0,  112.0,  121.0,  72.0, 30, 23.0, 26.2, 0.245]]) #765th patient
    array([0])

[ ] rfc.predict([[5.0,  175.0,  166.0,  72.0, 51, 19.0, 25.8, 0.587 ]]) #769th patient
    array([1])
```

Figure 8: Predicting Diabetes Mellitus

From the above figure 8, we can clearly see that the 1st, 4th and 769th patient has returned the predicted value as '1' which means he/she may suffer from diabetes mellitus whereas the 765th patient has predicted value as '0' which means this patient is not suffered from diabetes.

8 Results Analysis

After splitting the diabetes dataset into Training and Testing Test, at first, the research work implements all existing machine learning algorithms to training and testing test and produce the following results:

Table 3: Comparison of Existing Machine Learning Algorithms

	Accuracy (in %)	Precision (in %)	Recall (in %)	F1 Score (in %)
Bagging Classifier	88	79	81	80
XGBoost Classifier	88	78	83	81
AdaBoost Classifier	88	81	79	80
Random Forest	88	78	84	81
Gaussian Naïve Bayes	79	66	66	66
Kneighbours	88	85	74	79
LinearSVC	70	100	1	3
Logistic Regression	80	69	61	65
Decision Tree	83	71	74	73

Graphical representation of comparison of machine learning algorithms with respect to Accuracy score, precision score, recall and f1 sore.

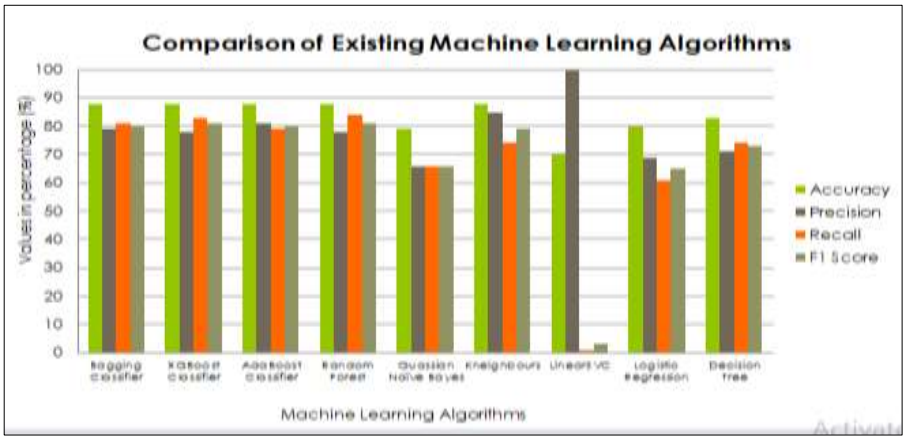


Figure 9: Comparison of Existing Machine Learning Algorithms with Accuracy, Precision, Recall and F1 score

The research work also shows the cross validation scores for base model. It is an important step in machine learning process and helps us to ensure that the model selected for deployment is robust and generalizes well to new data. The following figure 10 shows the cross validation result.

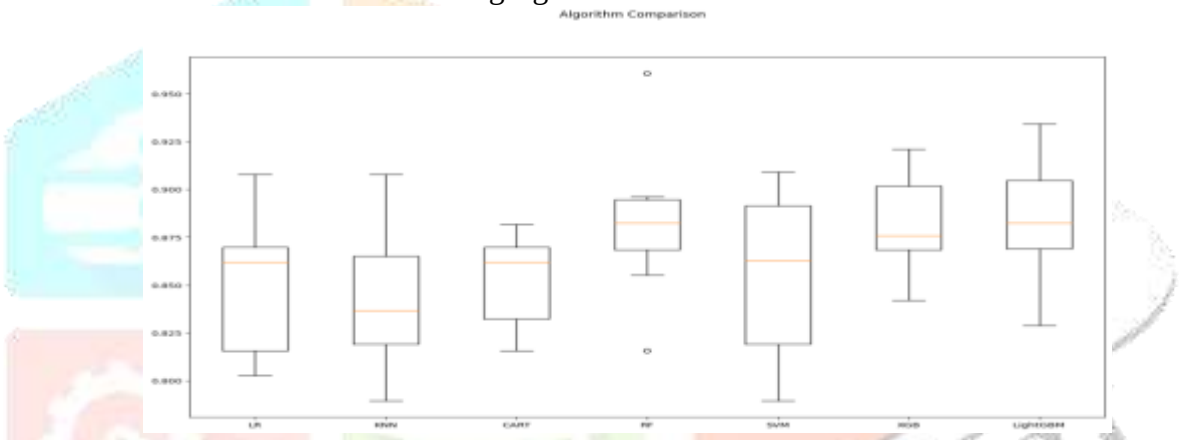


Figure 10: Comparison of Base model cross validation score

Then the research work installs the final model with XGBoosting algorithm, random forest algorithm and LightGBM algorithm. The following figure 11 shows the final model comparison of Hybrid XGBoosting Algorithm (Ensemble Gradient Boosting Algorithm), LightGBM Algorithm and Random Forest Algorithm for cross validation scores of all the instances.

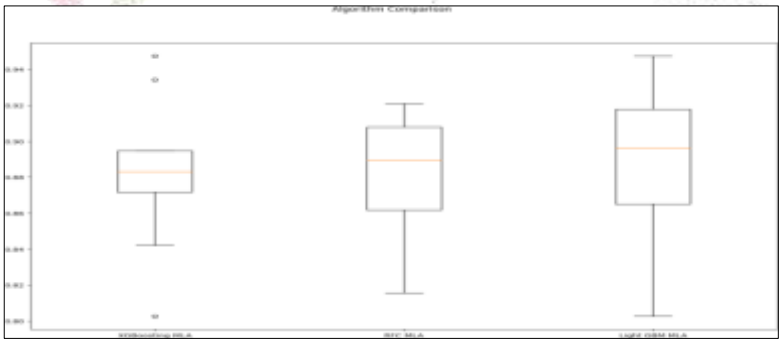


Figure 11: Comparison of final model cross validation scores

From the above figure 11, it is clearly visible that the Ensemble Gradient XG Boosting model gives the lowest cross validation scores as compared to others.

After obtaining optimized cross validation score, the research work implements the ensemble gradient boosting Machine Learning Algorithm along with K Nearest Neighbors classifier, Logistic Regression,

Gaussian Naïve Bayes and Linear SVC one by one to predict diabetes mellitus and produce the better accuracy score, precision score, recall score, AUC and ROC curve scores. The results are shown in following table.

Table 4: Results of Ensemble Gradient Boosting Classifier, KNN Classifier, Gaussian Naïve Bayes, Logistic Regression and Linear SVC Algorithm

Machine Learning Algorithms	Training Data Accuracy Scores	Testing Data Test Accuracy Scores	Precision Scores	Recall Scores	Area Under Curve Scores
GradientBoostingClassifier	0.9869	0.8821	0.786667	0.842857	0.871114
KNeighborsClassifier	0.9006	0.8603	0.756757	0.800000	0.843396
LogisticRegressionCV	0.8086	0.7948	0.707754	0.571429	0.732255
GaussianNB	0.8180	0.7904	0.657143	0.657143	0.753100
LinearSVC	0.8235	0.4760	0.362637	0.842857	0.606649

From the above table, it is clearly visible that with the help of ensemble technique, ensemble Gradient Boosting algorithm has produced best accuracy score(0.9869), precision(0.8821), recall(0.842857) and AUC score(0.871114) as compared to all other algorithms.

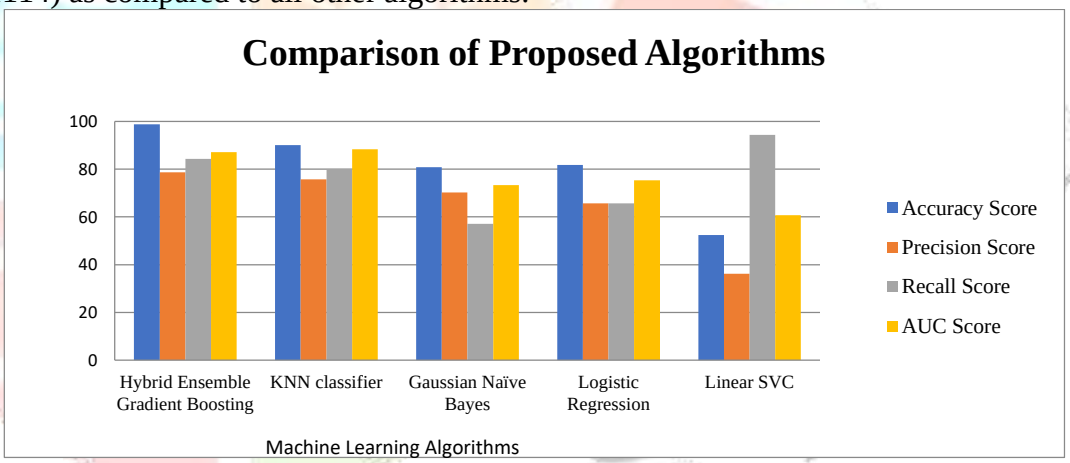


Figure 12: comparison of Ensemble Gradient Boosting Classifier, KNN Classifier, Gaussian Naïve Bayes, Logistic Regression CV and Linear SVC Algorithm with Accuracy, Precision, recall and AUC score for Trained Dataset

From the above figure 12, we can clearly see that the ensemble hybrid gradient boosting machine learning algorithm outperforms the rest of the algorithms.

After implementing all machine learning algorithms along with ensemble gradient boosting algorithm to produce better accuracy score, precision score, recall and AUC score, the research work gives better ROC curve scores which is shown in following figure.

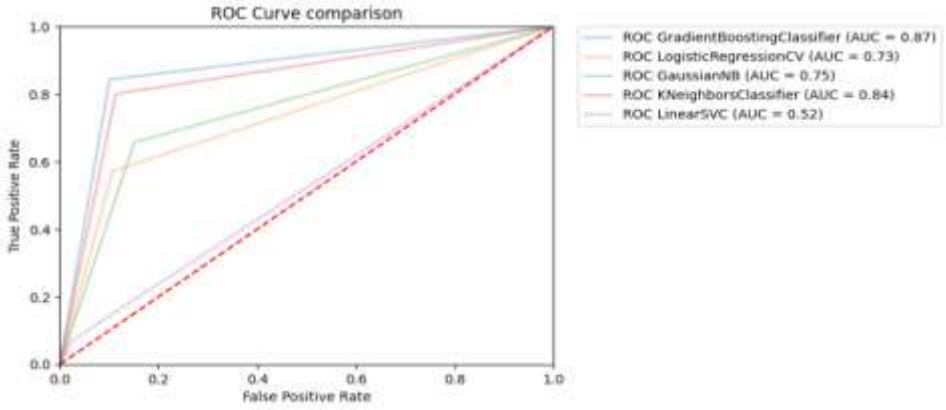


Figure 13: comparison of Ensemble Gradient Boosting Classifier, KNN Classifier, Gaussian Naïve Bayes, Logistic Regrssion CV and Linear SVC Algorithm with ROC curve

From the above figure 13, it is clearly visible that the Ensemble Gradient Boosting Classifier has produced maximum no. of ROC curve score i.e., 0.87, as compared to rest of the Machine Learning Algorithms which clearly means that the ensemble gradient boosting performs better prediction by distinguishing the instances falls under exact True Positive Rate and False Positive Rate. The following figure shows the comparison of all machine learning algorithms with proposed hybrid ensemble gradient boosting algorithm.

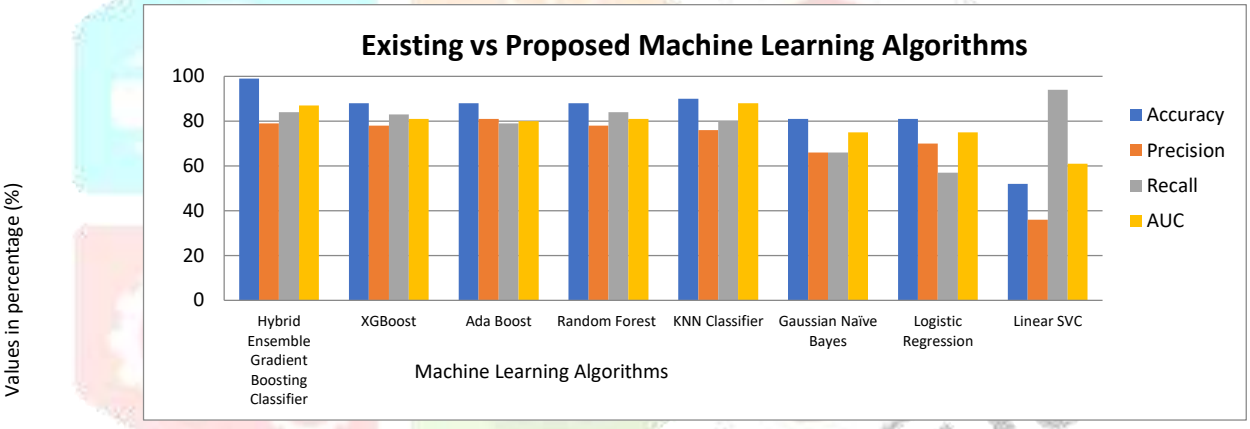


Figure 14: Comparison of all Machine Learning Algorithms with proposed Hybrid Ensemble Gradient Boosting Classifier

The above figure 14 shows the comparison of all Machine Learning Algorithms with proposed Hybrid Ensemble Gradient Boosting Classifier. Here, Hybrid ensemble Gradient Boosting Machine Learning Algorithm along with seven machine learning algorithms namely XGBoost classifier, AdaBoost classifier, Random Forest classifier, Gaussian Naïve Bayes algorithm, KNN Classifier algorithm, Logistic Regression and LinearSVC algorithm has been implemented and produce results with respect to their accuracy score, precision, recall and AUC.

From the above figure 14, it is clearly visible that Hybrid Ensemble Gradient Boosting algorithm has produced highest and better accuracy i.e., 99% whereas KNeighbor Classifier Algorithm has produced the next highest accuracy i.e., 90% followed by XGBoost classifier, AdaBoost Classifier and Random Forest Classifier with second highest accuracy i.e., 89%, Gaussian Naïve Bayes algorithm (81%), Logistic RegressionCV(82%) and LinearSVC with 52% accuracy. And it is also clearly visible that the highest correct predictions, recall and AUC scores has been obtained by hybrid ensemble gradient boosting algorithm.

4 Conclusion and Future Scope

The research work is carried out for diabetes patients of Andaman & Nicobar Islands. This research not only showcase exploratory data analysis features through graphical representations but also gives best accuracy, precision and recall score by using ensemble gradient boosting machine learning model. The research work uses several machine learning algorithms like Gaussian Naïve Bayes algorithm, KNN classifier, Logistic Regression and LinearSVC along with ensemble gradient boosting model to find out the better results. The research work also gives better cross validation score and produce better AUC and ROC curve scores using ensemble gradient boosting classifier algorithm. From the results and data analysis, it is finally concluded that the ensemble gradient boosting algorithms gives best results as compared to rest of the machine learning algorithms. So this algorithm can be helpful in further studies of healthcare sector for predicting any kind of chronic and dangerous disease like cancer, TB, Kidney disease, etc.

References

[1] Mohammad Ehsan et al. : Detection and prediction of diabetes using effective biomarkers, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, Volume 12, 2024 - Issue 1.

[2] Ashisha GR, Anitha Mary X and Mahimai Raja J: Classification of Diabetes Using Ensemble Machine Learning Techniques, Scalable Computing: Practice and Experience, Vol. 25 No. 4 (2024).

[3] P. Senthil Kumari: Predicting the Severity of Diabetes Using ECLAT Algorithm in Data Mining, Proceedings of the International Conference on Digital Transformation in Business: Navigating the New Frontiers Beyond Boundaries (DTBNNF 2024), Series:Advances in Economics, Business and Management Research, 2024.

[4] Praveena Nuthakki and T. Pavan Kumar: Machine learning-based early detection of diabetes risk factors for improved health management. Multimedia Tools Application (2024). https://doi.org/10.1007/s11042-024-18728-5 April 2024.

[5] Kirti Kangra and Jaswinder Singh: A genetic algorithm-based feature selection approach for diabetes prediction, IAES International Journal of Artificial Intelligence (IJ-AI), Vol.13(2), June 2024.

[6] Shahid Mohammad Ganie et. al.: An ensemble learning approach for diabetes prediction using boosting techniques, Front. Genet., 26 October 2023 Sec. Computational Genomics, Volume 14 - 2023 | https://doi.org/10.3389/fgene.2023.1252159 Joanish Muthu and S. Suriya : Type 2 Diabetes Prediction using K-Nearest Neighbor Algorithm, Journal of Trends in Computer Science and Smart Technology Vol.5(2), June 2023.

[7] Kirti Kangra and Jaswinder Singh: Comparative analysis of predictive machine learning algorithms for diabetes mellitus, Bulletin of Electrical Engineering and Informatics Vol 12, Issue 3, June 2023.

[8] Boon Feng Wee et.al. : Diabetes detection based on machine learning and deep learning approaches, Multimedia Tools and Applications, Vol. 83, Issue 5:Pp.1-33, August 2023.

[9] Liangjun Jiang et. al.: Diabetes risk prediction model based on community follow-up data using machine learning, Preventive Medicine Reports, Vol. 35, Issue 11, October 2023.

[10] Michael OnyemaEdehet.al.,: A Classification Algorithm-Based Hybrid Diabetes Prediction Model,

Front. Public Health, 31 March 2022 Sec. Digital Public Health, Volume 10 (2022).

[11] Sasmita Padhy et.al,:IoT based Hybrid Ensemble Machine Learning Model for Efficient Diabetes Mellitus Prediction, Computational Intelligence and Neuroscience special issue, Volume 2022.

[12] Rishab Bhotra: Diabetes Predictions using Machine Learning Algorithms, International Journal of Engineering Applied Sciences and Technology, Vol. 6. Issue 5, ISSN No. 2455-2143, Pp. 151 – 154 (2021).

[13] Akihiro Nomuraet. al.: Artificial Intelligence in Current Diabetes Management and Prediction, Springer Article number: 61 (2021).