Available online at http://scik.org
J. Math. Comput. Sci. 11 (2021), No. 4, 4008-4022
https://doi.org/10.28919/jmcs/5750
ISSN: 1927-5307

IMPROVISING HEALTHCARE DECISION MAKING BY EMPLOYING ENSEMBLE TECHNIQUE

NARENDRA KUMAR SHARMA[1,†,*], SHAHNAZ FATIMA[1], SWATI SAXENA[2]

[1]Amity Institute of Information Technology, Amity University Uttar Pradesh, Lucknow, India

[2]Dept. of Computer Application, Maharana Pratap Engineering College, Kanpur, Uttar Pradesh, India

Abstract: Today, World is facing an elevated threat from a multitude of diseases that are growing at an astronomical rate every day. The numbers of diseases detected by medical institutions are increasing every year. Early prediction or detection of any disease can help people cure it to the fullest. After the initial cure, whether disease will create further health issues in future is also the area of investigation. Thus, predicting disease is a more important task to help clinicians to provide effective treatment for people. In this paper, we combine several classification approaches to improve the accuracy of the classifier. We propose an iterative ensemble approach that constructs a powerful classifier by mixing manifold low-performance classifiers with the intention that a powerful classifier with high precision can be obtained. The dataset used in this work maintains approx 50 attributes of diabetic patients. We examine whether after the initial recovery the patient has health issue in future or not.

Keywords: Data Mining; Machine Learning; Healthcare; Decision Making.

2010 AMS Subject Classification: 90B50.

1. INTRODUCTION

Data Mining includes insights from a variety of disciplines including artificial neural networks,

*Corresponding author

E-mail address: narendrasharmauim@gmail.com

†Research Scholar

Received March 22, 2021

computational learning theory, pattern recognition, statistics, genetic algorithms and probabilistic modelling. Hence, it embraces a wide range of techniques which takes place whilst learning, for instance recognition methods such as kNN and instance-based learning, discriminant analysis, and Bays classifiers. Some problems arise in the study of patient data because these datasets are imperfect (lost parameter values), have inaccuracies (systematic or random noise in the data), sparseness (poor patient data are available and / or unrepresentable).

The increasing supremacy of machine learning in diagnosing disease and in arranging and categorizing health information will authorize general practitioner and accelerate decision making in the health centre. The healthcare system records huge voluminous of data on patient's details and it becomes a tedious and difficult task to analyse those data for the humans.

1.1 Machine Learning in Healthcare

Machine learning (ML) techniques help to build decision support models and gives explanation about the data. It provides an easy method for healthcare professionals to analyse the data and provide better diagnosis of the disease. The machine learning techniques used in the healthcare system requires the storage of adequate data and the authorization to use that data [2].

1.2 Ensemble Technique

Ensembles are composite models that combine a series of poorly performing classifiers to create an improved classifier. It returns the votes of the individual classifiers performing the majority of the votes and the final prediction label. Ensembles proffer higher precision than individual or basic classifiers. Ensemble approaches can be parallelized by assigning each base learner to a distinct machine. Ultimately, ensemble learning techniques are meta-algorithms that coalesce manifold ML techniques into a solitary prognostic model to perk up performance. Ensemble technique may employ bagging to diminish variance, boosting approach to trim down bias, and stacking approach to perk up prediction [3].

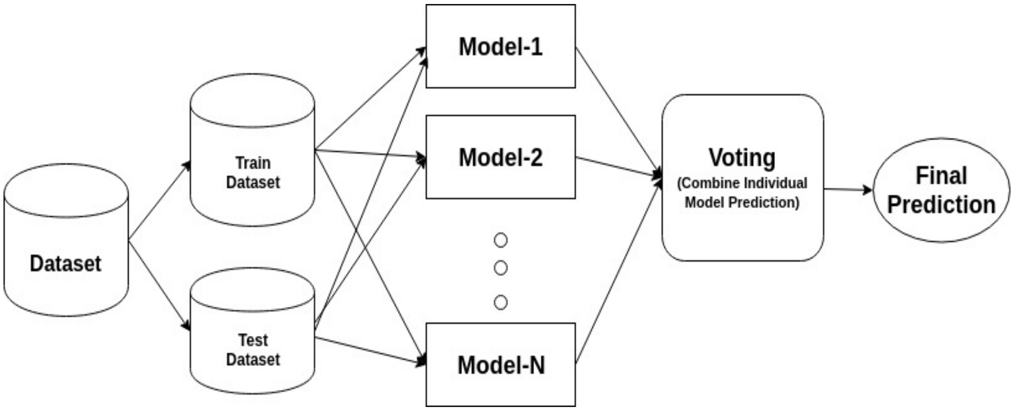

Figure 1: Ensemble Technique

Ensemble techniques can be divided into two groups depending on the category of base learners. Homogeneous ensemble techniques use base learners of the same kind for each iteration. The heterogeneous ensemble method utilizes diverse kinds of base learners for each iteration.

2. LITERATURE REVIEW

- Ioannis Kavakiotisa[6] explored the use of data mining approaches and ML in diabetes research. The author performed a detailed analysis of diverse techniques of data mining used for envisaging diabetic complications and their effectiveness in other genetic diseases.
- Simon Fong et al.[7]presented a data stream mining methodology to derive real-time decision rules for identifying blood-glucose reactions on patients based on the insulin levels. Decision trees are framed based on the present health conditions monitored from the patient continuously rather than using the history of records for decision making.
- Machine learning algorithms use the classifier system to solve health care problems. These are used to assist doctors in identifying and forecasting diseases in early stages Abdelhmid Salih et al.[8] But, as the medical information set is unorganized, heterogeneous with high dimensions and noise, it is challenging to extract data from medical information records. The idea is to choose an appropriate technique after analyzing all the available methods. There is a rise in the dependence of using medical data for diagnosing a disease.
- Jose Sanz et al.[9] adopted linguistic fuzzy rule based classifier to predict the risk that a patient likely to suffer heart disease. The adopted framework diagnoses and also provides brief interpretation on the decision and will enable the doctors to acquire information from the available history of patient records.
- Decision making process is to classify a new data by comparing its features with that of clustered data for which the classes are identified already. Rodrigo C Barros et al.[10] proposed automated decision tree induction framework derived to perform a particular method of classification in various domains of application. The authors performed experimentation in actual world gene expression datasets to evaluate effectiveness of proposed framework in classification task.

3. PROPOSED APPROACH

In the proposed ensemble boosting multiple classifiers are mingled to enhance the accuracy of classifiers. We combine classifiers with low performance to create a powerful classifier that produces a powerful classifier with high precision. The vital concept behind this proposal is to set classifier weights and train a sample of data at each iteration to guarantee perfect prophecy of anomalous observations. Any ML approach may be utilized as a base classifier as long as it admits training set weights. Proposed approach is expected to satisfy following two conditions:

1. The training of classifier must be made interactively using a variety of examples of weighted learning.
2. At each iteration, the classifier tries to present the best for these examples while diminishing learning errors.

Following are the steps for its working:

1. First, a training subset is arbitrarily picked.
2. Train the proposed ML model iteratively by choosing a training set based on precise predictions of the final training.
3. Assign superior weights to misclassified cases so that those cases are more likely to be classified in the subsequent iteration.
4. Also assign weights to the classifiers trained in every iteration, depending on the correctness of classifier. The better classifier gets more weight.
5. We repeat the same procedure unless entire training data matches without errors or unless the utmost number of specified evaluators is reached.
6. To categorize, "vote" on all the learning algorithms that you created.

Model hyperparameters are characteristics of the model that are external to the model and cannot be estimated from the data. The hyperparameter values must be set before starting the training process. Starting our approach, we set some important hyperparameters, such as:

- base_estimator: weak learner used to train models. DecisionTreeClassifier is used as the default weedy learner.
- n_estimators: the number of weak learners who train iteratively.
- Learning_rate: promotes weight gain for weak learners. Use 1 as default.

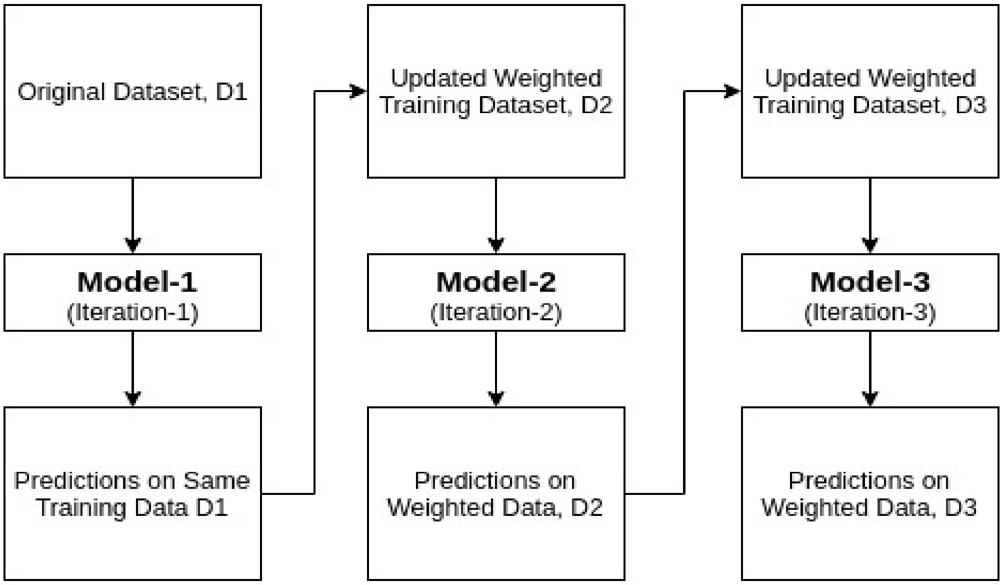
Original Dataset, D1
Updated Weighted Training Dataset, D2
Updated Weighted Training Dataset, D3
Model-1 (Iteration-1)
Model-2 (Iteration-2)
Model-3 (Iteration-3)
Predictions on Same Training Data D1
Predictions on Weighted Data, D2
Predictions on Weighted Data, D3

Figure 2: Proposed iterative approach

3.1 Parameter Tuning

Grid search algorithms and Bayesian optimization algorithms are generally used in machine learning to tune hyperparameters. When hyperparameter adjustments are made, the grid search algorithm may look like a local optimization rather than a global optimization, but a Bayesian optimization algorithm can provide global optimizations. This study used grid search.

4. APPLICATION IN HEALTHCARE DECISION MAKING

The rate of readmission is now deemed as a measure of the excellence of a hospital and also has a negative impact on the cost of treatment. It also throws the light on the immunity system of patients. Readmission to hospital for diabetics is costly because hospitals may face fines if readmission rates are higher than expected and reflect a poor health care system. It is also useful for the patients while deciding the hospital for admission. For these reasons, it is important for hospitals to pay more attention to reducing readmission rates. In this paper we identify the key factors influencing readmission of diabetic patients and predict the likelihood of patient readmission.

5. METHODOLOGY

In this research work, the dataset chosen is that available on the UCI website which contains the patient data for the 10 years for 130 hospitals. We implemented proposed approach in Python using different libraries like scikit-learn, seaborn, matplotlib etc. The flow of implementation is depicted in figure 3.

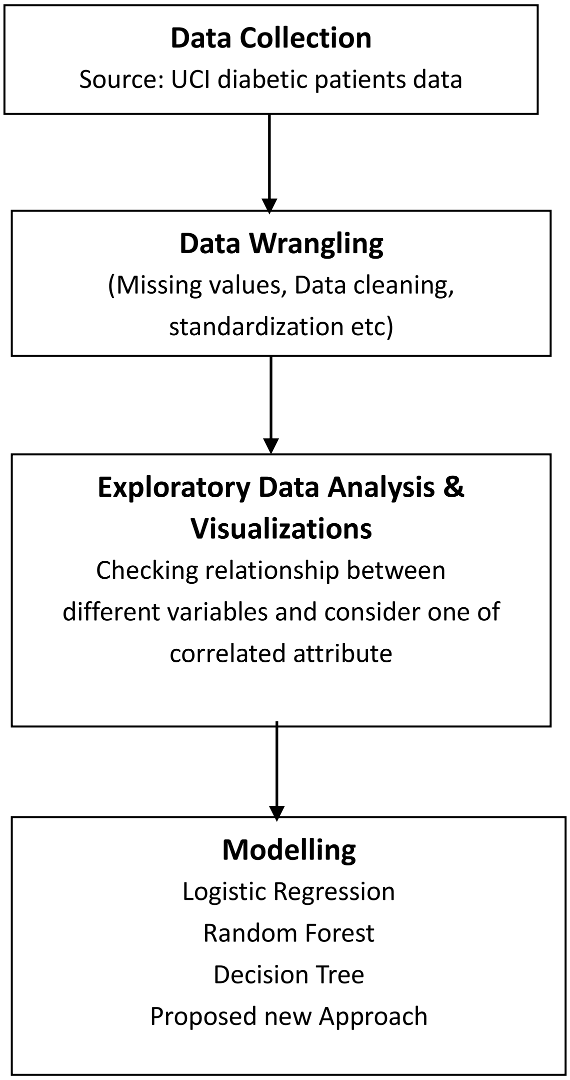
Data Collection
Source: UCI diabetic patients data
Data Wrangling
(Missing values, Data cleaning, standardization etc)
Exploratory Data Analysis & Visualizations
Checking relationship between different variables and consider one of correlated attribute
Modelling
Logistic Regression
Random Forest
Decision Tree
Proposed new Approach

Figure 3: Flow of Implementation

5.1 Data Set

The main data source is the UCI website, which represents the Clinical Care of patient in an integrated distribution network of 130 hospitals. It includes over 50 functions that represent patient and hospital results. There are about 100,000 entries in total. This dataset contains information about appointments that satisfy following conditions:

1. There is Hospitalization
2. The patient has diabetes

3. Duration of stay from one day to fourteen days.
4. During the admission time, laboratory tests were carried out.
5. The medicine was administered during the admission. Data includes number of patients, race, gender, age, type of hospitalization, length of stay, hospital specialization, number of laboratory tests performed, HbA1c test results, diagnosis, number of drugs, diabetes medications, number of outpatients, etc. like hospitalization and emergency visits in the previous year.

5.2 Exploratory Data Analysis

We performed exploratory data analysis to get a basic sense about the data. Check relationship between different variables to understand the data and if there is a strong correlation between two variables then we can consider one of them. When we analyze dataset we find that there is no problem of multi-collinearity.

5.3 Modelling

Logistic Regression, Random Forest, Decision Tree and Proposed new Approach were applied considering following features of dataset:

> Weight, Admission type, Medical specialty, Admission source, Number of lab procedures, Discharge disposition, Number of medications, Time in hospital, Number of outpatient visits, Number of procedures, Number of emergency visits, Number of inpatient visits, Glucose serum test result, Diagnosis 1, Diagnosis 2, Diagnosis 3, Number of diagnoses, Change of medications, HbA1c test result, Diabetes medications, 24 features for medications (For the generic names: repaglinide, metformin, glyburide, insulin nateglinide, chlorpropamide, acetohexamide, glimepiride, tolazamide, tolbutamide, glipizide, glimepiride-pioglitazone, miglitol, pioglitazone, rosiglitazone, examide, acarbose, troglitazone, sitagliptin, ,metformin-rosiglitazone, glyburide-metformin, glipizide-metformin, and metformin-pioglitazone) and Readmitted.

System was trained to identify the target class in Readmitted feature. Calculate the accuracy, precision, support, recall, f1-score, macro and weighted average for judging model performance.

Accuracy: It is just the percentage of predictions that were made correctly and can be calculated as:

$$Accuracy = \frac{Number\_of\_Correct\_\mathrm{Pr}edictions}{Total\_\mathrm{Pr}edictions}$$

It can be computed in terms of TP, TN, FP and FN[16] as follows:

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

Precision: It is the division of accurately envisaged positive interpretations to the entirety of envisaged positive interpretations.

$$\Pr ecision = \frac{TP}{FP + TP}$$

Recall: It is the division of accurately envisaged positive interpretations to all the interpretations in original class labelled as yes.

$$\operatorname{Re} call = \frac{TP}{FN + TP}$$

F1 Score: It is the harmonic average of recall as well as precision.

$$\text{F1 score} = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \times \frac{\text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$

Macro-average: It is computed separately for every performance measure.
Weighted-average: It is weighted average. Its calculation for F1 Score is as:

Weighted Average (F1 Score) = $\frac{\sum_{i=1}^{n} F_i \times N_i}{N}$

where F_i is F1 Score for ith class, N_i is the count of instances in ith class, N is the total count of instances and n is the count of classes in the dataset.

6. EXPERIMENTAL EVALUATION

Before applying ML and proposed approach, we make the data fit as follows

- Cleaning the data, replacing the nullvalues in numeric data by 0 and object data by unknown,
- Substituting 0 and unknown,
- Encoding the data,
- Drop the irrelevant columns
- Normalization of the data,

- Divide the data into training and validation data sets. Training data will contain 80% of data and validation will contain remaining 20%

Categorical variables have been explored in figure 4.

We applied the proposed approach and the existing approach to determine the confusion matrix[11]. A confusion matrix is a two-dimensional array that compares predicted categorical labels with actual labels. For binary classification, these are categories of TP, TN, FP and FN. The visualizations of the confusion matrix for logistic regression [12], random forest [13], decision tree [14] and the proposed new approach are shown in Figures 5-8.

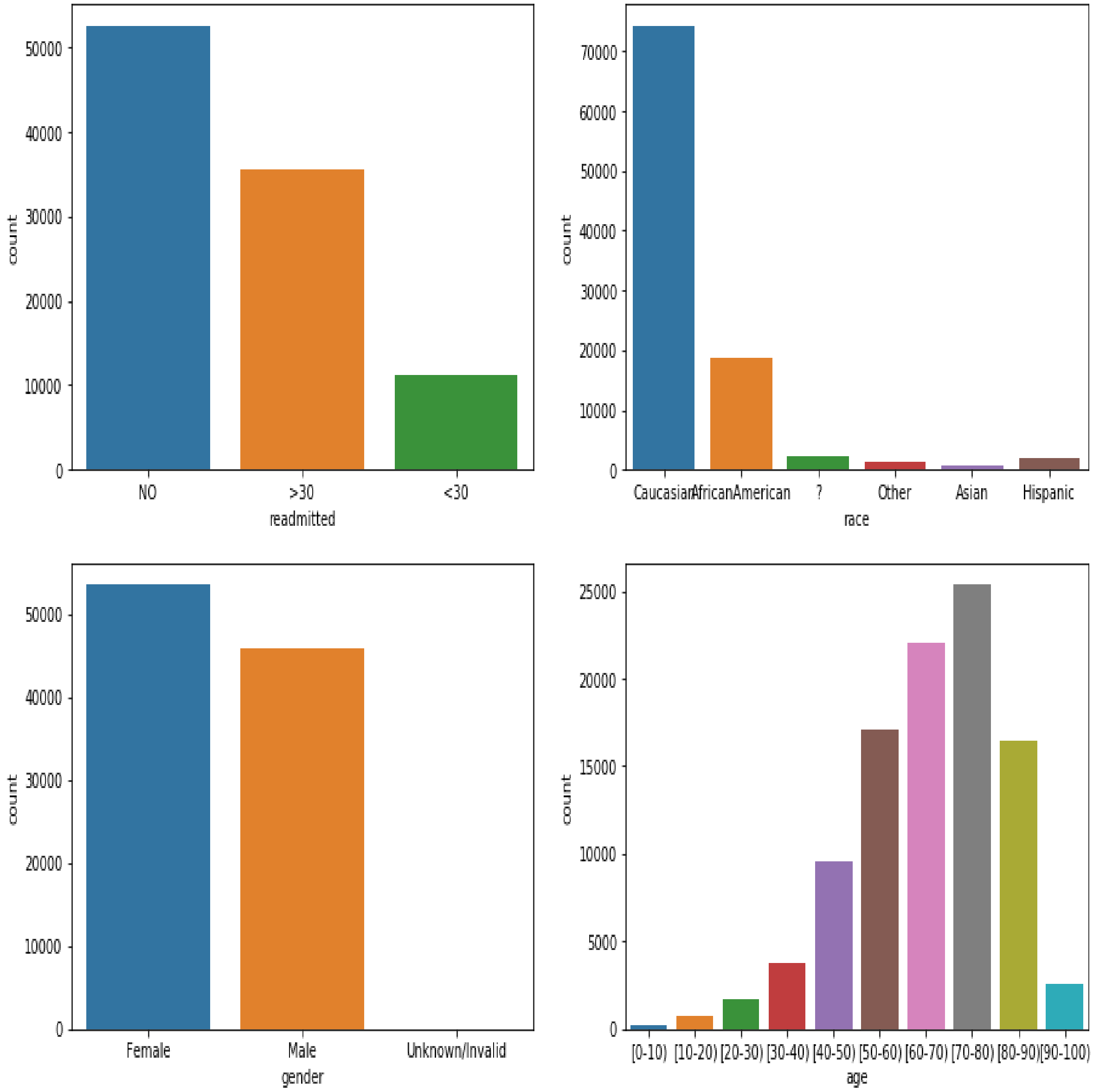

Figure 4: Categorical variables in diabetic patients dataset

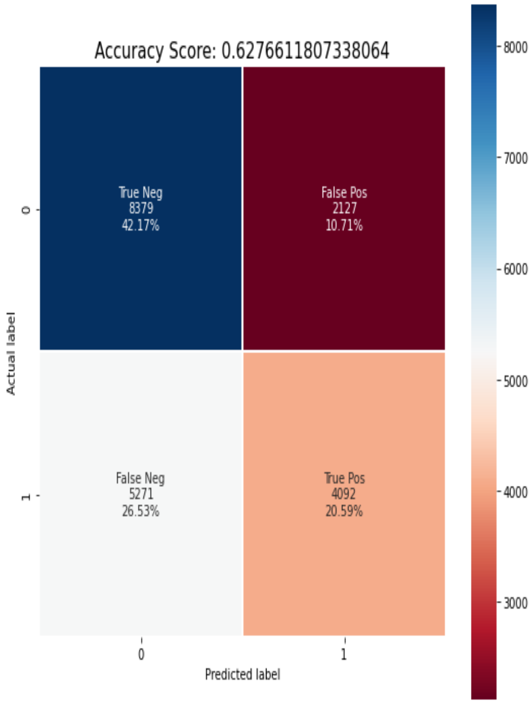

Figure 5: Confusion matrix visualization for Logistic Regression

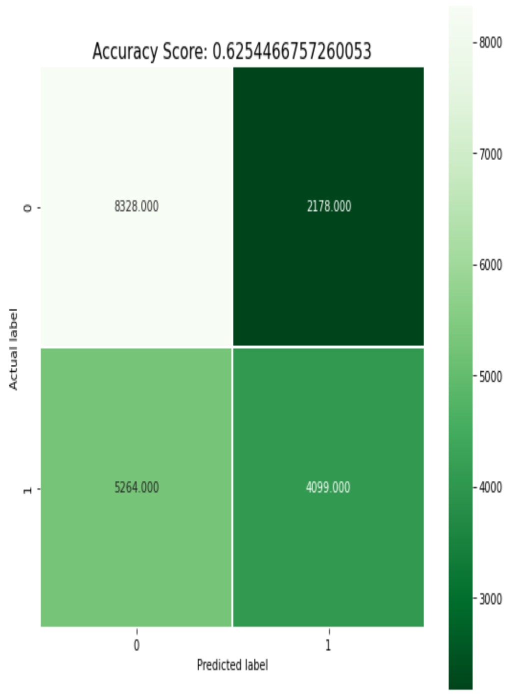

Figure 7: Confusion matrix visualization for decision tree

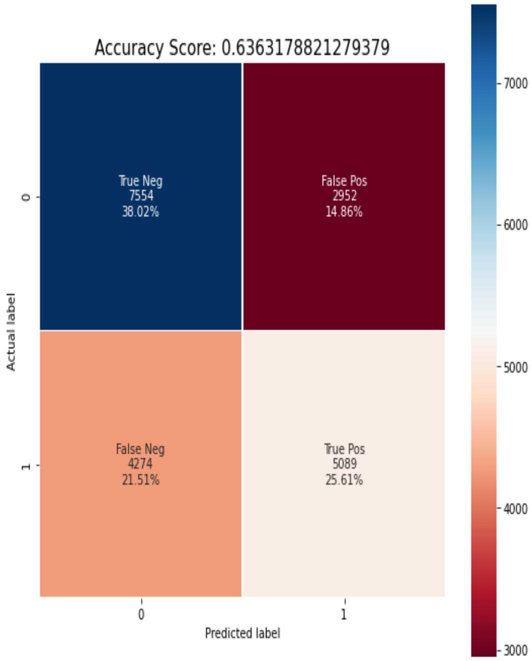

Figure 6: Confusion matrix of Random Forest

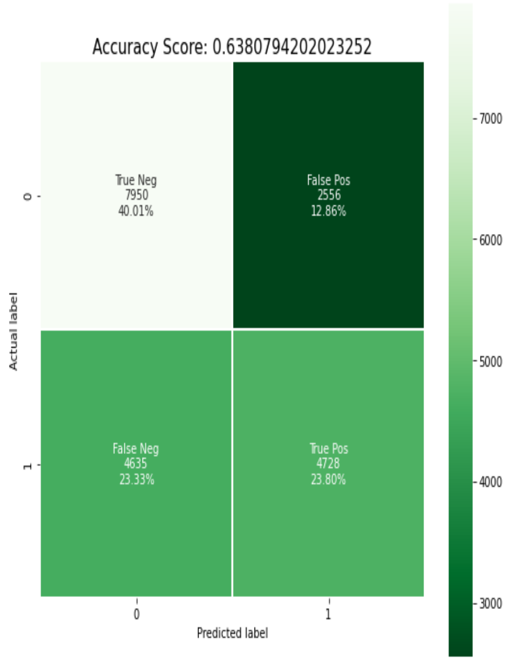

Figure 8: Confusion matrix visualization for proposed approach

7. COMPARISON

The results are expressed using four score precision, recall, f1-score, support for each as below.

Logistic Regression

	precision	recall	f1-score	support
NO	0.61	0.80	0.69	10506
YES	0.66	0.44	0.53	9363

accuracy			0.6277	19869
macro avg	0.64	0.62	0.61	19869
weighted avg	0.63	0.63	0.61	19869

Random Forest

	precision	recall	f1-score	support
NO	0.64	0.72	0.68	10506
YES	0.63	0.54	0.58	9363

accuracy			0.6363	19869
macro avg	0.64	0.63	0.63	19869
weighted avg	0.64	0.64	0.63	19869

Decision Tree

	precision	recall	f1-score	support
NO	0.62	0.74	0.68	10506
YES	0.63	0.49	0.55	9363

accuracy			0.6236	19869
macro avg	0.62	0.62	0.61	19869
weighted avg	0.62	0.62	0.62	19869

Proposed Approach

	precision	recall	f1-score	support
NO	0.63	0.76	0.69	10506
YES	0.65	0.50	0.57	9363

accuracy			0.6381	19869
macro avg	0.64	0.63	0.63	19869
weighted avg	0.64	0.64	0.63	19869

The performance of classifier at all thresholds can be graphically shown by ROC Curve (Receiver Performance Curve)[15]. ROC curve displays 2 parameters: the rate of TPs and the rate of FPs. ROC curve between different models and the proposed approach shown in Figure 9.

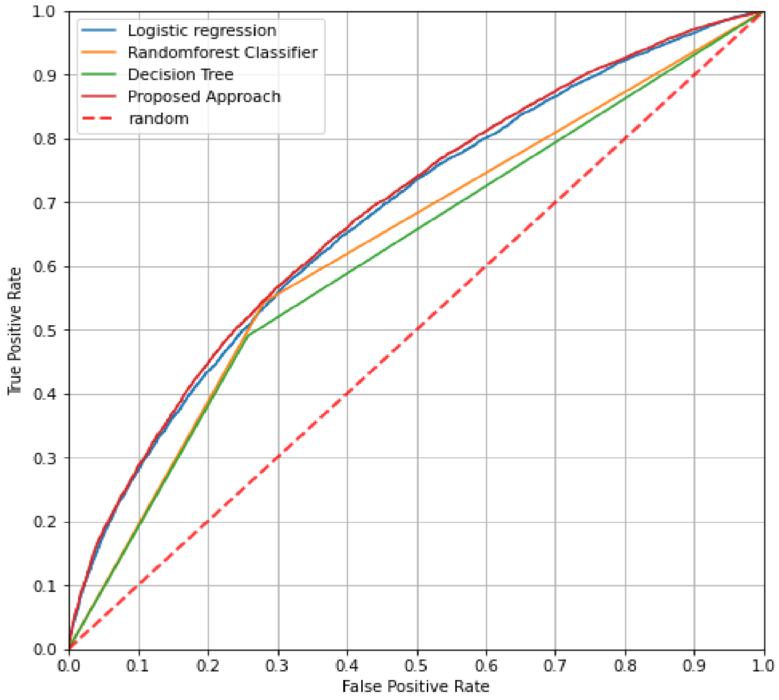

Figure 9: **ROC curve** of proposed and other approaches

Comparison of accuracy of approaches is depicted in figure10.

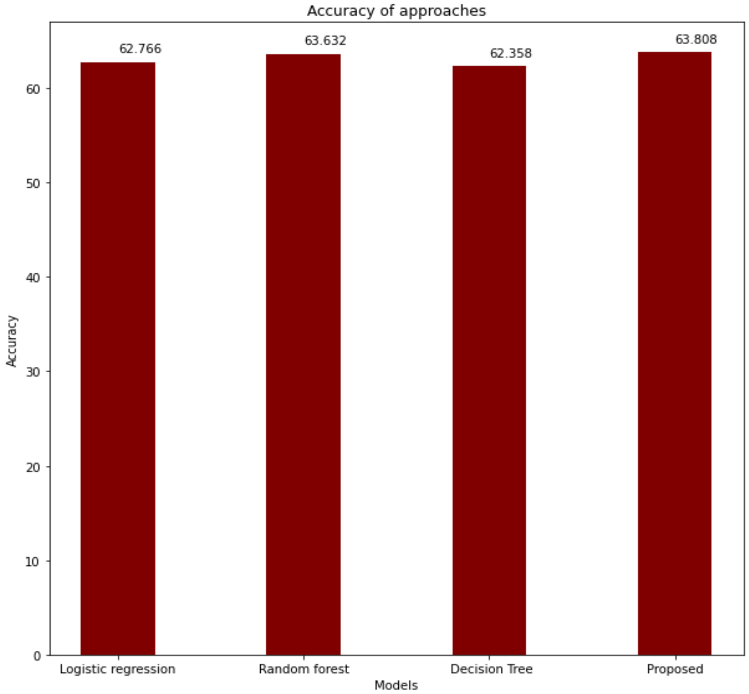

Figure 10: Accuracy of approaches

Thus we can simply come to the conclusion that the proposed approach clearly identifies that in future health issues of the patient which is justified by readmission of patient in the hospital.

8. Conclusion

An earnest effort has been made to develop disease diagnosis system that facilitates identification and prediction of the disease from the patient data. The reliance of data mining and ML techniques in classification of disease has been acknowledged and the algorithms are implemented for assisting health care professionals. In this paper we proposed an iterative ensemble method which constructs a burly classifier by mingling numerous scantily performing classifiers with the intention that we get elevated accuracy. After comparing the accuracy of proposed approach with other existing approaches, we conclude that the accuracy of proposed approach is slightly better than others. The same approach will work in any case of any disease in the dataset is balanced.

Conflict of Interests

The author(s) declare that there is no conflict of interests.

REFERENCES

[1] P.S. Mung, S. Phyu, Effective analytics on healthcare big data using ensemble learning, in: 2020 IEEE Conference on Computer Applications (ICCA), IEEE, Yangon, Myanmar, 2020: pp. 1–4.

[2] J. de la Torre, J. Marin, S. Ilarri, J.J. Marin, Applying machine learning for healthcare: a case study on cervical pain assessment with motion capture, Appl. Sci. 10 (2020), 5942.

[3] D. Chicco, G. Jurman, An ensemble learning approach for enhanced classification of patients with hepatitis and cirrhosis, IEEE Access. 9 (2021), 24485–24498.

[4] P. Arjun, J. Verma, Methods for detection of diabetes mellitus using machine learning techniques, J. Multidiscip. Eng. Sci. Technol. 7(11) (2020), 12948-12956.

[5] C. Krittanawong, H.U.H. Virk, S. Bangalore, et al. Machine learning prediction in cardiovascular diseases: a meta-analysis, Sci Rep. 10 (2020), 16057.

[6] I. Kavakiotis, O. Tsave, A. Salifoglou, et al. Machine learning and data mining methods in diabetes research, Comput. Struct. Biotechnol. J. 15 (2017), 104–116.

[7] S. Fong, J. Fiaidhi, S. Mohammed, L.A.M. Moutinho, Managing diabetes therapy through datastream mining, IT Professional, 19 (2017), 50–57.

[8] A.S.M. Salih, A. Abraham, Novel ensemble decision support and health care monitoring system, J. Netw. Innov. Comput. 2 (2014), 041-051.

[9] J.A. Sanz, M. Galar, A. Jurio, A. Brugos, M. Pagola, H. Bustince, Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system, Appl. Soft Comput. 20 (2014), 103–111.

[10] R.C. Barros, M.P. Basgalupp, A.A. Freitas, A.C.P.L.F. de Carvalho, Evolutionary design of decision-tree algorithms tailored to microarray gene expression data sets, IEEE Trans. Evol. Computat. 18 (2014), 873–892.

[11] M. Hasnain, M.F. Pasha, I. Ghani, M. Imran, M.Y. Alzahrani, R. Budiarto, Evaluating trust prediction and confusion matrix measures for web services ranking, IEEE Access. 8 (2020), 90847–90861.

[12] S. Bhandari, A.S. Shaktawat, A. Tak, et al. Logistic regression analysis to predict mortality risk in COVID-19 patients from routine hematologic parameters, Ibnosina J. Med. Biomed. Sci. 12(2) (2020), 123-129.

[13] C. Iwendi, A.K. Bashir, A. Peshkar, et al. COVID-19 patient health prediction using boosted random forest algorithm, Front. Public Health. 8 (2020), 357.

[14] M.M. Ghiasi, S. Zendehboudi, A.A. Mohsenipour, Decision tree-based diagnosis of coronary artery disease: CART model, Computer Meth. Programs Biomed. 192 (2020), 105400.

[15] A.C.J.W. Janssens, ROC curves for clinical prediction models part 2. The ROC plot: the picture that could be worth a 1000 words, J. Clinic. Epidemiol. 126 (2020), 217–219.

[16] V.M. Patro, M.R. Patra, Augmenting weighted average with confusion matrix to enhance classification accuracy, Trans. Mach. Learn. Artif. Intell. 2(4) (2016), 77-91.