International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

Optimization Accuracy for Diabetes Diagnosis and Prediction using Machine Learning Technique

Deepti Shikha Ojha
Department of Computer Science and Engineering
NIRT, Bhopal
Dr. Devendra Kumar Bajpai
Department of Computer Science and Engineering
NIRT, Bhopal

Abstract

The goal of this research is to improve the overall disease prediction accuracy by analyzing the automatic prediction and recommendation of diabetes disease from the electronic health record diabetes dataset. Diabetes data is acquired from patients and are processed utilizing optimal artificial intelligence techniques during the diabetes data recognition process. This research integrated machine learning based approaches to predict diabetes disease features such as: SVM, DT, RF, LR, K-NN, NB and GB. The proposed GB model is proposed to apply diabetes diagnosis which is single class and multiclass classification problems. In the future, we shall incorporate an auto feature selection method to design the crossed features and select the features for the prediction and classification model. Subsequently, the Inclination helping calculation gives the best exactness to Diabetes finding than the past calculation.

Keywords: - Diabetes Disease, Machine Learning, Gradient Boosting, Accuracy

INTRODUCTION

Growth of data has tremendously increased due to the advancement in social networking and innovative gadget delivery in the competitive digital world. The healthcare industries play a vital role in the growth of data consistently under the category of patients care in terms of food, nutrition, physical activity and various other monitoring requirements [1, 2]. This tremendous growth in data causes difficulties in handling data [3]. Emerge of multiple new techniques can majorly supports to predict and minimize the cost for predicting the prevalence of various long-lasting diseases. These techniques are supportable in a way to continuous monitoring of health treatments and their benefits over the patients" health improvement. Because of the advancement of technology the gathering and monitoring the improvement is highly possible which essentially supports to manage various health needs [4]. Behavioral data obtained from these devices helps in collecting healthcare data like current health condition and recognize efficient tool for further analysis on disease prediction.

The improved performance of these analytic tools plays major role in reduction of expenses related to healthcare and its progress monitoring process. Similarly the application of analytic tools and techniques widespread to do analysis oven psychological health, ecological health long-lasting and communicable illness prediction, quality enhancement of treatment, accidents forecast and analyze its severity. Apart from these promises, there are few issues related to data security and its manipulation on time. The healthcare data arises from various sources right from person physical activity to data generated from healthcare centers under various treatments. There are multi variant data populated in each healthcare centers indirectly support for mining multiple predictions over diseases [5].

HEALTHCARE DATA

Healthcare data are ever growing data and the advancement of technology stimulates the growth further. The volume of the data is exponentially increased and it becomes very difficult to handle. It basically the hospitals maintain their patients Electronic Medical Records in a healthcare repository.

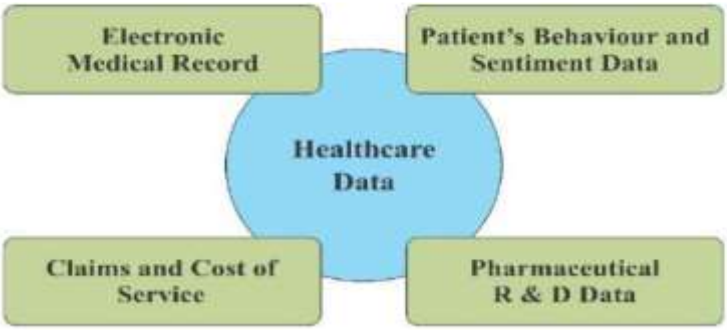

Electronic Medical Record
Patient's Behaviour and Sentiment Data
Healthcare Data
Claims and Cost of Service
Pharmaceutical R & D Data

Figure 1: Sources of Healthcare Data

There are other three main sources of healthcare data such as diseases and their related symptoms supplied by pharmacological centers, the data related to patient's mannerism and also various emotional data of each patient collected from healthcare centers as shown in Figure 1.

The data growth in medicines and clinics comprises numerous categories .It is often fully related to medics and patients. Further, the data in medication are produced from various other clinical activities. The healthcare data focuses mainly on the collection of various medical equipment's of different domains of healthcare.

Based on the demand and the availability of data from patient records, survey and database are maintained in hospitals, which include details from a patient admission to discharge summary. Handling these diversified data sources are challenging one and it is necessary to analyze and abstract them properly for further prediction of diseases. There are huge amount of healthcare data produced every day with an extent of 90% of the overall population and the gadgets in use. The population of healthcare data involves the application of a variety of medical equipment's which are embedded with advanced techniques [6, 7].

Analyzing Healthcare Data

The analysis of healthcare data majorly provides better prediction result on the prevalence of disease and the ratio of disease impact upon the society. Though there are numerous sources of data, the healthcare industry has been the important source of formal and informal data. As per the report of IDC, the global healthcare data has been increased from 500 petabytes to 25,000 petabytes by 2020. The data from laboratory test, genomic data, data from various sensorial sources wearable over the body etc. provides sufficient space for disease related analyses and prediction. By the way the process over the healthcare data helps in secure life span of people and helps the public to be aware of prevalent diseases, impact of those diseases upon human being, what are the safety measures to be followed in their day-to-day life to plan their future expenses related to healthcare. Basically, the result of each examination helps to enhance the decision making ability of each individual [8]. Below Figure 2 shows the flow of healthcare data analytics. It involves sources of data, process of data analytics and its resultant cost effective decisions. There are four basic types of data analysis, such as descriptive analysis, diagnostic analysis, predictive analysis and prescriptive analysis.

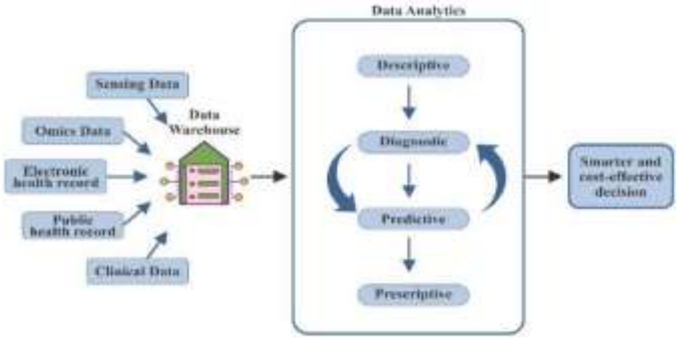
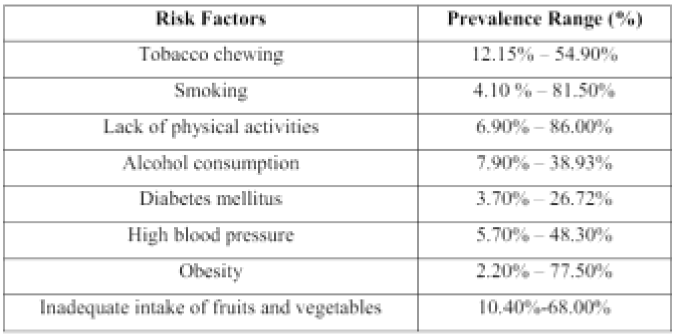


Figure 2: Flow of Healthcare Data Analytics

RISK FACTORS OF DIABETES DISEASE

Following a global declaration against NCDs, a national program for Non-Communicable Diseases (NCDs), was launched in India a decade ago. Being a diverse country with a population of 17% of the world, the pattern, distribution of diseases and their determination are very diverse, factors that affect the selection and delivery of evidence for controlled interventions.

Table 1: Pervasiveness of Various Risk Factors



Risk Factors	Prevalence Range (%)
Tobacco chewing	12.15% – 54.90%
Smoking	4.10 % – 81.50%
Lack of physical activities	6.90% – 86.00%
Alcohol consumption	7.90% – 38.93%
Diabetes mellitus	3.70% – 26.72%
High blood pressure	5.70% – 48.30%
Obesity	2.20% – 77.50%
Inadequate intake of fruits and vegetables	10.40%-68.00%

As a country of diversity, the same size is consistent with all the unequal principles in the implementation of interventions in the Indian provinces. Chewing tobacco, smoking and drinking alcohol are found to range from12.15-54.9%, 4.1-81.5% and 7.9-38.93% respectively. The prevalence of obesity varied between 2.4-33.1% and 2.2-77.5% while the lack of moderate physical activity varies from 6.9-86.0%. Diabetes as a risk factor ranges from 3.7-26.72% while the estimated high BP varies between 5.7- 48.3% as shown in Table 1.

According to the survey, the urban areas in Tamil Nadu (13.7%) and Jharkhand (13.5%) have the maximum rates of diabetes. Rural areas record the least rates that are only 3% in Jharkhand and 6.5% in Maharashtra. The statistics enforces special attention on diabetes and its related complications are essential. The proposed research focuses on Type-2 Diabetes and the health complications related to it.

Basically diabetes is metabolic illness due to over utilized and underutilized level of blood glucose of every individual. An existence of ratio of diabetes increases in low and middle income countries. The untreated diabetes causes high risk of difficulties such as eye impairment, kidney impairment, impairment of nervous system, hearing loss, Alzheimer's and vascular disease. There are numerous risk factors related to diabetes. But there is lack of awareness among population and are leading their life as insecure. The foremost work on these analysis is to track from the factors which causes diabetes [9, 10]. So the present study focuses on diabetes and the factors related to it.

PROPSOED METHODOLOGY

Making a classification model out of a dataset with labelled classes and some features, like a dependent binary variable and an independent variable, is the main objective of machine learning techniques. The majority of the GB machine algorithms' workflow is composed of the training and dataset validation phases. Using the training dataset, the method adjusts the prediction model to reduce error in the output results. Due to the separation of the training dataset and the validation dataset, the learning algorithm was developed independently. The main objective of this measure is to determine the endpoint of the training method in order to stabilize the trained model's accuracy against overfitting.

Additionally, GB classifier gives better outcomes as far as expectation; nonetheless, it needs more opportunity to prepare in the cycle interaction. Further Improvised-GB gives improved preparing model and exactness, additionally it centers around issues such as tree learning, enhanced tree learning aids in discovering the best split, thus to accomplish that we build up a specific algorithm which is provided later in a similar segment.

Algorithm for Proposed Methodology:-

```
Step 1: Start
Step 2: Import Library: define all loading dataset, visualization, data preprocessing, data splitting, confusion matrix, machine learning and accuracy library
Step 3: Upload Dataset (files .upload command)
Step 4: Stored in a Pandas Data frame (pd.read_csv command)
Step 5: Pre-processing
Step 6: Variable (input variable)
    (6.1) if
    (6.2) (Input variable >= 2)
    (6.3) end if;
Step 7: Classes; 0 for not disease and 1 for disease
Step 8: Applied Machine learning Technique
Step 9: Create Confusion Matrix
Step 10: Threshold (T), maximum value of T is 100
    (10.1) if
```

```
        (10.2) (Accuracy >= T)
        (10.3)  end if;
Step 11: Calculate Parameter
Step 12: End
```

SIMULATION RESULTS

So, the accuracy can be measured according to Eq. 5.1

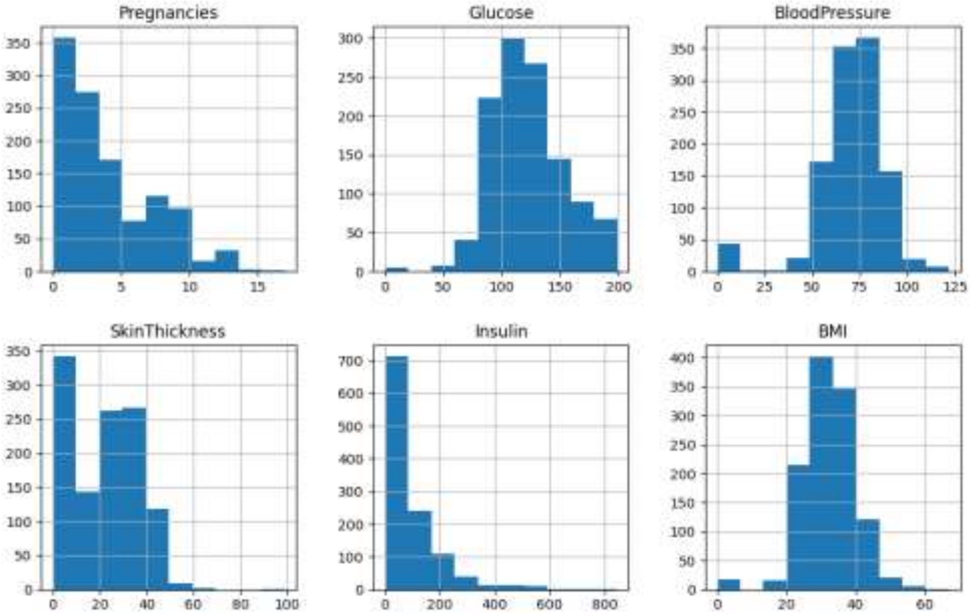
$$Accurancy = \frac{TN + TP}{TN + TP + FN + FP} \qquad (1)$$

For a diabetes classification problem, its measures include Precision-Recall and accuracy. The formula to derive these measures is given in Eq. 2 and Eq. 3.

$$\mathrm{Pr}ecision = \frac{TP}{TP + FP} \qquad (2)$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \qquad (3)$$

Figure 3 shows the histogram of attributes and the range of dataset attributes and code used to create it.



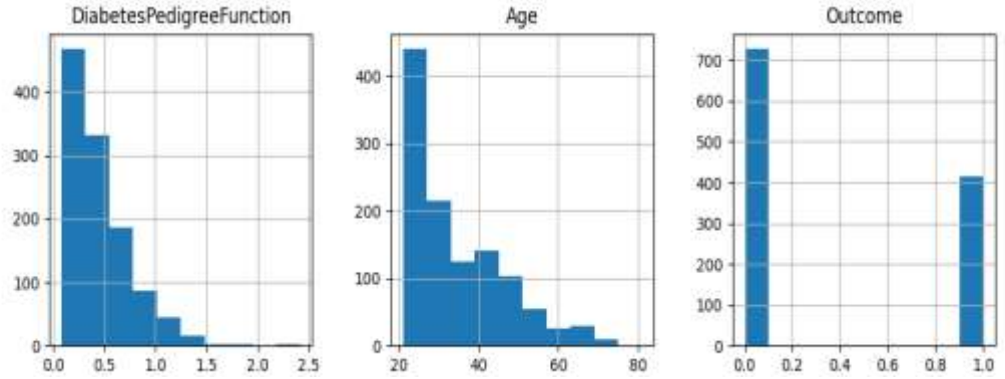


Figure 3: Histogram of Dataset

Figures 5 and 6 show the status of Diabetes health, ranging from healthy to severely unhealthy. Blue bar represents Diabetes disease, and the red bar represents not Diabetes disease

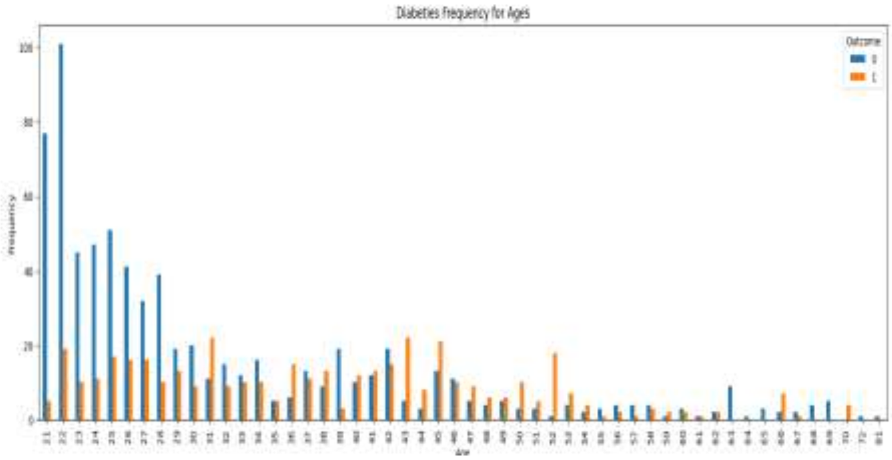


Figure 4: Bar Plot of the Number of Diabetes Frequency for Ages

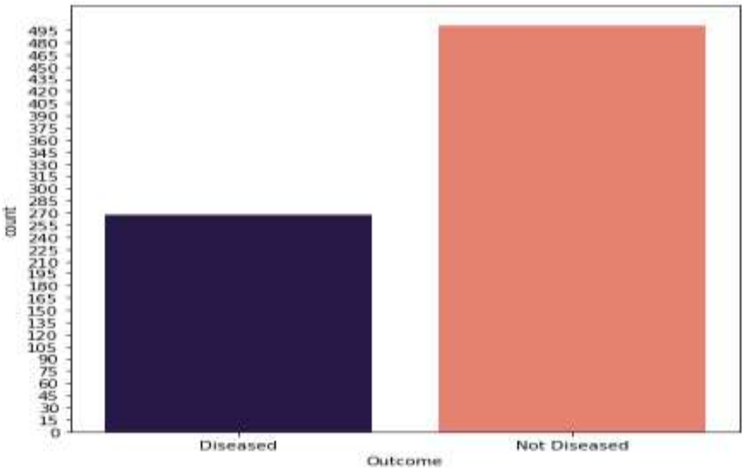


Figure 5: Bar Plot According to Outcomes

Table II represents the accuracy for different ML classifier with Previous Isfafuzzaman Tasin et al. [1]. GB classifier is best accuracy compared to Previous Isfafuzzaman Tasin et al. [1].

Bar Graph of the previous and proposed Algorithm for Accuracy in Diabetes Dataset is representing in fig. 6.

Table II: Comparison Result for Accuracy

Techniques	**Previous Isfafuzzaman Tasin et al. [1]**	**Proposed Algorithm**
LR	75.00%	77.41%
GNB	79.00%	75.95%
RFC	76.00%	82.69%
K-NN	73.00%	73.90%
DT	72.00%	83.57%
SVM	78.00%	78.29%
XGBoost Classifier	81.00%	94.18%

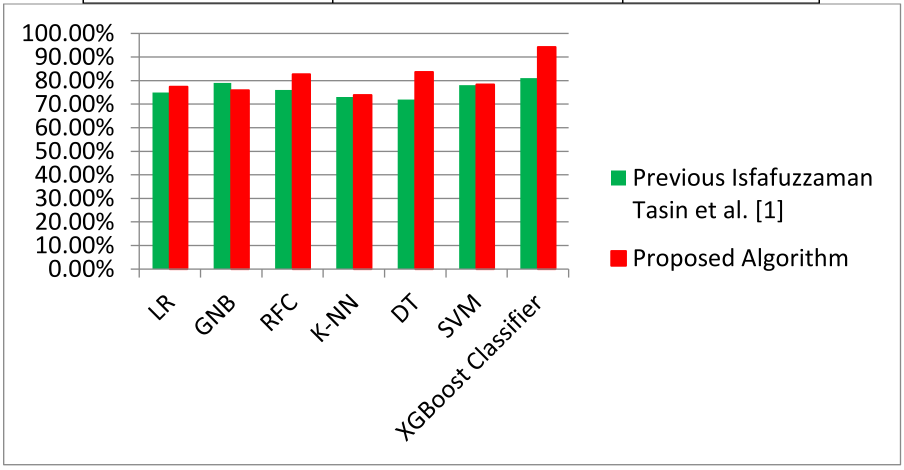


Figure 6: Bar Graph of the Previous and Proposed Algorithm for Accuracy

CONCLUSION

Diabetes is a chronic metabolic disorder with more rising prevalence among the people worldwide. Aiming to improve the treatment of people with diabetes, ML techniques play a vital role with its advancement in analysis and promising results. In the present research work, a systematic literature survey is conducted and applied suitable ML techniques for analyzing the prevalence of diabetes and its related complications.

REFRENCES

[1] Isfafuzzaman Tasin, Tansin Ullah Nabil, Sanjida Islam, Riasat Khan, “Diabetes prediction using machine learning and explainable AI techniques”, Healthcare Technology Letters, pp. 01-10, Wiley 2022.

[2] Olisah, C.C., Smith, L., Smith, M., “Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective”, Comput. Methods Programs Biomed., Vol. 20, pp. 1–12, 2022.

[3] Deberneh, H.M., Kim, I., “Prediction of type 2 diabetes based on machine learning algorithm”, Int. J. Environ. Res. Public Health, Vol. 18, pp. 1–14, 2021.

[4] Nikos Fazakis, Otilia Kocsis, Elias Dritsas, Sotiris Alexiou, Nikos Fakotakis, and Konstantinos Moustakas, “Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction”, IEEE Access 2021.

[5] Naveen Kishore G, V.Rajesh, A.Vamsi Akki Reddy, K.Sumedh, T.Rajesh Sai Reddy, “Prediction of Diabetes using Machine Learning Classification Algorithms”, International Journal of Scientific & Technology Research, Vol. 9, No. 01, 2020.

[6] Chatrati, S.P., Hossain, G., Goyal, A., “Smart home health monitoring system for predicting type 2 diabetes and hypertension”, J. King Saud Univ. Comput. Inf. Sci., Vol. 34, No. 3, pp. 862–870, 2020.

[7] Hasan, M.K., Alam, M.A., Das, D., Hossain, E., Hasan, M., “Diabetes prediction using ensembling of different machine learning classifiers”, IEEE Access, Vol. 8, pp. 76516–76531, 2020.

[8] Cervantes, J., García-Lamont, F., Rodríguez, L., Lopez-Chau, A., “A comprehensive survey on support vector machine classification: Applications, challenges and trends”, Neurocomputing, Vol. 408, pp. 189–215, 2020.

[9] Pranto, B., “Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh”, Information Vol. 11, pp. 1–20, 2020.

[10] Mohan, N., Jain, V., “Performance analysis of support vector machine in diabetes prediction”, In: International Conference on Electronics, Communication and Aerospace Technology, pp. 1–3, 2020.

[11] Muhammad Azeem Sarwar, 2Nasir Kamal, 3Wajeeha Hamid,4Munam Ali Shah, “Prediction of Diabetes Using Machine Learning Algorithms in Healthcare”, 24th

International Conference on Automation & Computing, Newcastle University, Newcastle upon Tyne, UK, 6-7 September 2019.

[12] Rao G.A., Syamala K., Kishore P.V.V., Sastry A.S.C.S. ., "Deep convolutional neural networks for sign language recognition", International Journal of Engineering and Technology(UAE) ,Vol: 7, Issue 5, pp: 62-70, 2018.

[13] Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju and Hua Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques", Springer, 2018.

[14] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," Neuro computing, vol. 237, pp. 350–361, May 2017.