The Diabetes Oracle: Insights from Machine Learning and Predictive Analytics

K.P. HARIPRIYA [0000-0002-7324-5418]
Department of Computer Science, Periyar University, Salem-11, Tamil Nadu, India.
priya22prakasam@gmail.com

H. HANNAH INBARANI * [0000-0002-2956-3507]
Department of Computer Science, Periyar University, Salem-11, Tamil Nadu, India.
hhinba@periyaruniversity.ac.in

***Abstract*- Diabetes is a significant health disorder with potentially severe implications for daily life, increasing the risk of various complications. According to a statistical report from 2045, an alarming estimate of around 135 million individuals may be affected by diabetes, underscoring the magnitude of the issue. Early prediction of diabetes based on symptoms is crucial for effective management. To this end, numerous computational algorithms are employed for early disease diagnosis. The classification of diabetes is done using machine learning techniques such as Naïve Bayes (NB), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), Ada Boosting (AB), K-Nearest Neighbors (KNN), Gradient Boosting (GB), and Logistic Regression (LR). Through evaluation measures, these methods are systematically compared. Experimental findings reveal that the Gradient Boosting and Ada Boosting methods achieve the highest accuracy of 98%, surpassing other techniques.**

Keywords: Diabetes Prediction, Machine Learning Algorithms, Ensemble Methods.

1. INTRODUCTION

Classification plays a vital role in the computational environment, especially in disease prediction. 2019 saw 463 million people worldwide suffer from diabetes, accounting for 9.3% of the total. Increases of 10.2% (578 million) by 2030 and 10.9% (700 million) by 2045 are predicted. 10.8% is the higher frequency in urban regions compared to 7.2% in rural areas, and 10.4% is the higher prevalence in high-income nations compared to 4.0% in low-income countries. It is concerning to note that 50.1% of people with diabetes do not know they have the disease. Additionally, impaired glucose tolerance affected an estimated 7.5% (374 million) of the global population in 2019, with projections indicating increases to 8.0% (454 million) by 2030 and 8.6% (548 million) by 2045 [Saeedi et al., 2019]. Figure 1 explains the statistical information about the diabetes and prediabetes population in 2022.

with the help of this survey is important. Nowadays, diabetes is a common issue, from babies to adults. It affects both genders and causes other issues in the human body. Diabetes [Mayo Clinic, 2024] can manifest in various ways. Watch out for things like always feeling thirsty, needing to go to the bathroom a lot, and losing weight without trying. If you notice ketones in your urine or feel tired and irritable, it could be a sign. Blurry vision, slow-healing sores, and frequent infections are also common. Recognizing these signs early and getting help can make a big difference in managing diabetes and staying healthy. Recognition of these blood tests is needed. To predict diabetes [Haripriya et al., 2022], the author uses different machine learning algorithms, such as NB, SVM, RF, DT, LR, KNN, AB, and GB. Ada Boosting and Gradient Boosting methods produce better results when compared to other algorithms. The AB and GB methods produce good results when compared to other algorithms.

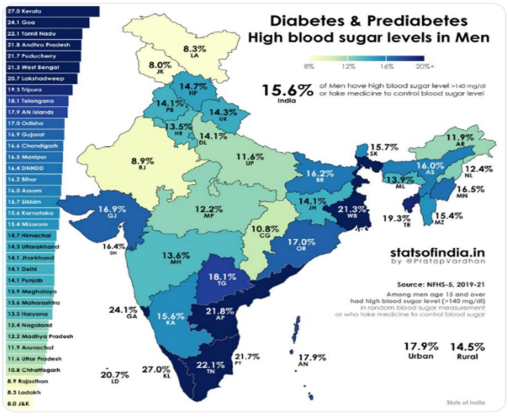


Figure 1. Statistical information on diabetes in 2022

The following is a description of the subsequent portions of this work: A thorough analysis of the perspectives of several writers is provided in Section 2. Section 3 elaborates on the proposed methodology and introduces innovative algorithms. In Section 4, meticulous attention is given to the evaluation parameters for segmentation along with their respective accuracy values. Finally, Section 5 presents the findings of the study, including an analysis of the most effective methods for the datasets utilized.

2. RELATED WORKS

This section mainly focuses on different author perceptions. According to the author [Mujumdar et al., 2019], predict diabetes using a machine learning method. The pipeline method with the machine learning algorithm is the suggested work that the author presented using two distinct datasets for the prediction: the PIMA Diabetes Dataset and the Diabetes Dataset. In that, the Ada Boosting classifier yields 98%, respectively. Using the diabetes dataset from Many machine learning methods are used by Kaggle [Rani et al., 2020]. These algorithms include random forests, logistic regression, decision trees, and support vector machines. Among them, DT yields good results with a 99% success rate. The machine learning algorithm-based ensemble methods are carried out [Hasan et al., 2020]. The author performed a K-fold cross-validation approach and compared it with several ensemble methods using the PIMA diabetic datasets to make their prediction. The author [Sonar et al., 2019] speaks about using a machine learning algorithm to predict diabetes. The features are taken out of the data before the algorithms are used. Ultimately, the data is divided using the 70–30 approaches, after which NB, SVM, DT, and ANN are employed. It has been investigated by Vijayan et al. (2015) whether diabetes and high plasma sugar levels are related. They suggested several AI-powered systems that use SVM, ANN, decision trees, and Naive Bayes classifiers to anticipate and diagnose diabetes. According to Pal et al. (2017), one of the main causes of visual impairment in diabetics is diabetic retinopathy (DR). As part of their study, they looked at how various machine learning algorithms performed and confirmed their effectiveness with a specific dataset.

3. RESEARCH METHODOLOGY

The train-test split, several machine learning methods, and data preprocessing are the key topics covered in this section. The paper's workflow is described in Figure 2. Before the raw data was entered into the classification framework, it underwent pre-processing. The entire data is split 80/20 between training and testing after pre-processing. Currently, diabetes data is classified using several machine-learning methods. Lastly, the accuracy and other metrics are computed using evaluation metrics to determine the extent to which the approach works better.

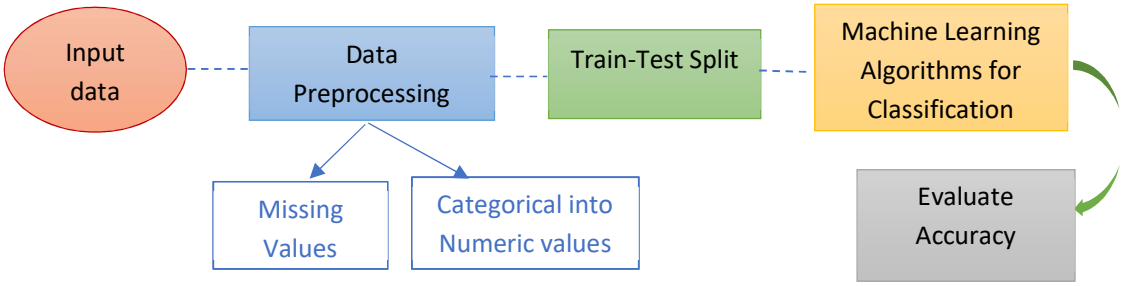


Figure 2. The workflow of this paper

3.1 Data Acquisition

The data must be cleansed before being fed into the model. Data preparation is the term for this procedure. Taking care of the missing data initially during the data preprocessing steps. The median of the data fills in the gaps in the data. Managing the category data is the second step. Since a machine can only comprehend binary information, One Hot Encoding helps transform category data into numerical data.

3.2 Different Algorithms

The data are divided into training and testing groups before being fed into machine learning algorithms; an 80/20 split is the most common. Currently, one lakh data points are divided into 20,000 data points for testing, with the remaining data points being used to train the model. Different machine learning algorithms are used for this work, such as DT, KNN, LR, NB, SVM, AB, GB, and RF methods. All the algorithms mentioned above fall under the category of supervised learning techniques. KNN, as described in [Azam et al., 2020], operates as a classification method. Its functionality relies on the K-component, which determines the nearest neighbors within the dataset. It is possible to generate forecasts by utilizing the K-component. The reference [Sailasya et al., 2021] discusses logistic regression for estimating the likelihood of binary output variables with a value of 0 or 1. In this work, the Decision Tree methodology, which is also mentioned in [Sailasya et al., 2021], is used for classification. It uses a tree-like structure and continuously splits data according to the entropy value. Random Forest, likewise cited in [Sailasya et al., 2021], consists of multiple independent decision trees trained on random subsets of data. Through a "voting" mechanism, the final prediction is determined, wherein each decision tree's class contributes to the outcome, with the most frequent class being selected as the final prediction. SVM is widely acknowledged as one of the best techniques for resolving classification, according to Soofi et al. (2017). In a Support Vector Machine (SVM), the support vector indicates the data point that is closest to the selection surface. It is especially good at binary classification and is frequently used when training data can be divided linearly. The last two approaches belong to ensemble algorithms. AdaBoost trains several inferior models independently, as referenced in [Pandey et al., 2016], and their combined predictions create the final prediction. Gradient boosting, as described by Cahyana et al. (2019), on the other hand, begins with a single leaf that has weights for every attribute. It initiates predictions of continuous values by taking the average value as the first guess. Notably, Gradient boosting and Ada boosting consistently achieve the highest accuracy compared to other methods.

4. OBSERVATIONS AND DISCUSSION

4.1 Description of the data

Based on [Mohammed Mustafa 2023] from Kaggle, the benchmark dataset for diabetes prediction is retrieved. It comprises a total of 101,000 data points stored in CSV format, featuring nine properties. These characteristics include blood pressure, history of heart disease, smoking, age, gender, Body Mass Index (BMI), HbA1c level, blood sugar level, and diabetes. The diabetes class label contains two categories: 0 denotes the absence of diabetes, while 1 signifies its presence. Notably, gender and smoking history are categorical variables, whereas the remaining properties are numerical.

4.2 Performance Metrics

The graphical representation of the table serves to evaluate the classification performance, providing a summary of how various models perform with the diabetes dataset [Khalif et al., 2024]. Within the confusion matrix, two labels are present: one for the actual class and the other for the predicted class. This matrix encompasses four variables: TN, TP, FP, and FN. Using metrics such as F1-Score, recall, accuracy, and precision, machine learning techniques are contrasted [Sharmila et al., 2021]. These metrics are computed using formulas outlined by [Ji et al., 2021], and they are derived from the confusion matrix.

$$Accuracy = \frac{Quantity\ of\ accurately\ identified\ pixels}{Total\ pixels\ count\ for\ the\ image} \quad (1)$$

$$Precision = \frac{TP}{(TP + FP)} \quad (2)$$

$$Recall = \frac{TP}{(TP + FN)} \quad (3)$$

$$F1 - Score = \frac{2(Precision * Recall)}{(Precision + Recall)} \quad (4)$$

The confusion matrix comprises four distinct instances: TN (true negatives) for correctly categorized negative instances, TP (true positives) for correctly categorized positive instances, False positives, or FPs, are positive cases that are erroneously classified, and false negatives, or FNs, are cases that are wrongly classified as negative. Several metrics are calculated using these variables, such as recall, accuracy, precision, and F1-score, which are listed in equations 1 to 4. Table 1 provides a breakdown of the evaluation metrics for different machine learning algorithms. SVM, DT, LR, AB, GB, RF, NB, and KNN are among the algorithms examined. These metrics aid in making predictions and selecting the most effective method, as depicted in the visual representation in Figure 2.

Measures	Methods							
	SVM	**DT**	**RF**	**AB**	**GB**	**LR**	**NB**	**KNN**
Accuracy	95.88	95.22	96.99	97.21	97.24	95.92	85.8	95.44
Precision	88.91	71.44	94.45	96.82	98.73	88.04	35.31	89.17
Recall	59.13	73.24	68.79	69.61	68.54	60.36	79.63	53.04
F1-Score	71.03	73.33	79.61	80.99	88.88	71.62	48.92	66.52

Table 1. Evaluation metrics results of different methods

Machine learning techniques come with a range of hyperparameters, which are detailed in Table 2. These hyperparameters were utilized throughout the prediction process. Figure 3 illustrates the various machine learning techniques employed for predicting diabetes, showcasing eight distinct approaches. The AB and GB algorithms stand out among these methods due to their better performance than the rest.

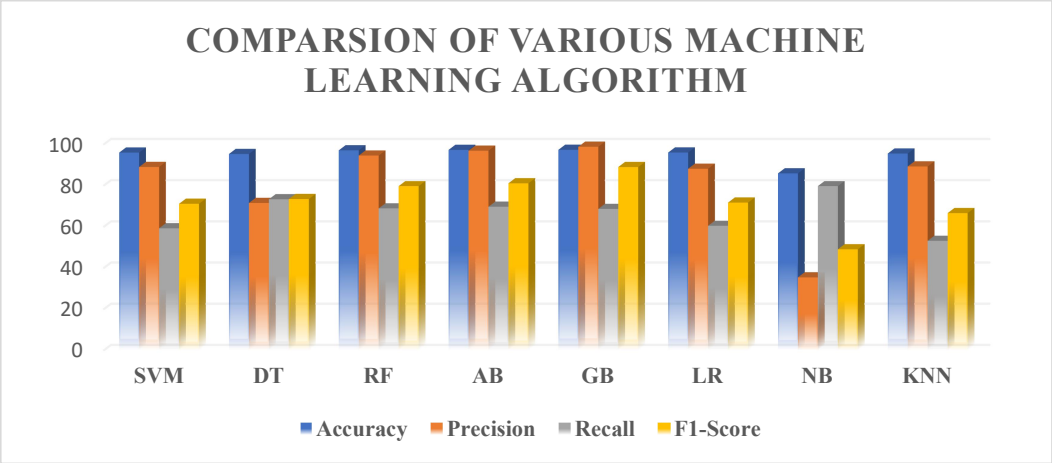


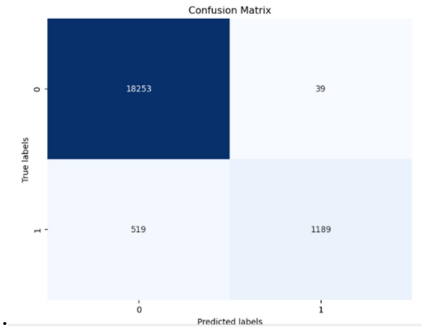
Figure 3. Different Algorithms for Diabetes Prediction

Table 2. Execution of the Techniques with Different Parameters

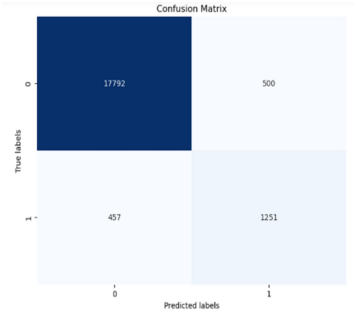
Methods	Accuracy	Precision	Recall	F1-Score
Multinomial Naive Bayes	89.93	38.91	31.44	34.78

Bernoulli Naive Bayes	90.99	36.77	07.73	12.77
SVM (RBF)	94.65	1.00	37.30	54.33
SVM(Sigmoid)	83.83	0	0	0

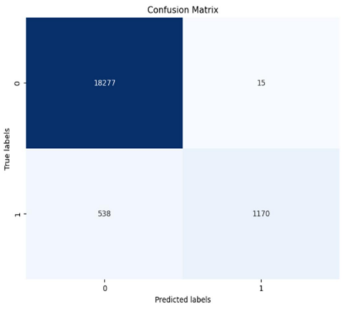
Table 2 presents hyperparameters for the different techniques utilized. Various Naive Bayes methods, such as Multinomial, Bernoulli, and Gaussian, are employed. SVM employs multiple kernels, including RBF, linear, and sigmoid. Notably, because there are no false negatives with the sigmoid approach, precision, recall, and f1-score are all shown as 0. Consequently, they are represented as 0 in Table 2. Figure 4's confusion matrix shows how different machine learning algorithm approaches perform. Some techniques demonstrate high true positives, while very few generate incorrect forecasts. Ultimately, Ada Boosting and Gradient Boosting stand out for their superior performance in true positive predictions compared to other techniques.



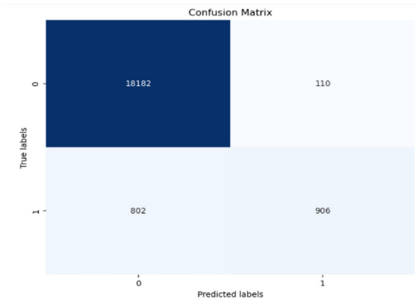
(A) Ada Boosting



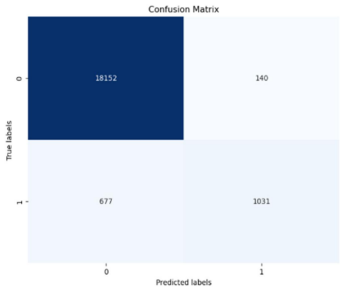
(b) Decision Tree



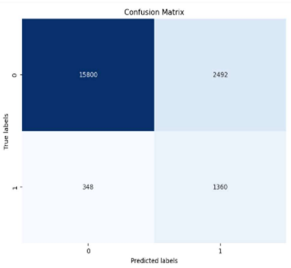
(c) Gradient Boosting



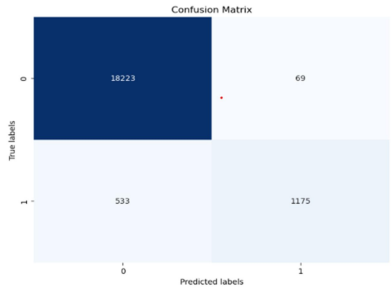
(d) KNN



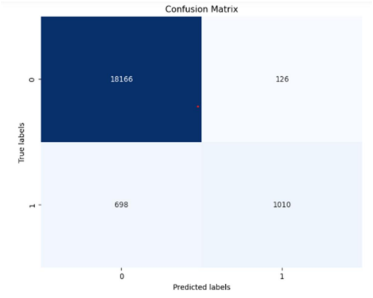
(e) LR



(f) NB



(g) RF



(h) SVM

Figure 4. Confusion Matrix for Different Methods

5. CONCLUSION

In the real world, detecting diabetes in the human body is crucial due to its potential association with heart attack risks. Among individuals aged 45 and above, some are diagnosed with diabetes, while others experience heart-related issues, and some remain unaffected. Various machine learning methods are employed in this endeavor, with Ada Boosting and Gradient Boosting algorithms demonstrating notable efficacy in diabetes prediction. Despite their effectiveness, these algorithms still exhibit misclassification of values. Looking ahead, integrating deep learning techniques or optimization algorithms holds promise for reducing such misclassifications. Initial assessments indicate that employing deep learning techniques on suitable datasets could yield favorable outcomes.

Abbreviations

SVM – Support Vector Machine
RF – Random Forest
AB – Ada Boosting
GB – Gradient Boosting
NB – Naïve Bayes
LR – Logistic Regression
DT – Decision Tree
TN – True Negative
TP – True Positive
ANN- Artificial Neural Networks

REFERENCES

Azam, M. S., Rahman, A., Iqbal, S. H. S., & Ahmed, M. T. (2020). Prediction of liver diseases by using a few machine learning-based approaches. Aust. J. Eng. Innov. Technol, 2(5), 85-90.

Cahyana, N., Khomsah, S., & Aribowo, A. S. (2019, October). Improving imbalanced dataset classification using oversampling and gradient boosting. In 2019 5th International Conference on Science in Information Technology (ICSITech) (pp. 217-222). IEEE.

Diabetes - Symptoms and causes - Mayo Clinic. (2024b, March 27). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444

Diabetes prediction dataset. (2023b, April 8). Kaggle. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

Haripriya, K. P., & Inbarani, H. H. (2022, December). Performance analysis of machine learning classification approaches for monkeypox disease prediction. In *2022 6th International Conference on Electronics, Communication and Aerospace Technology* (pp. 1045-1050). IEEE.

Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516-76531.

Ji, D., Zhang, Z., Zhao, Y., & Zhao, Q. (2021). Research on the classification of COVID-19 chest X-ray image modal feature fusion based on deep learning. *Journal of Healthcare Engineering*, *2021*.

Khalif, K. M. N. K., Chaw Seng, W., Gegov, A., Bakar, A. S. A., & Shahrul, N. A. (2024). Integrated Generative Adversarial Networks and Deep Convolutional Neural Networks for Image Data Classification: A Case Study for COVID-19. Information, 15(1), 58.

Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. Procedia Computer Science, 165, 292-299.

Pal, R., Poray, J., & Sen, M. (2017, May). Application of machine learning algorithms on diabetic retinopathy. In 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 2046-2051). IEEE.

Pandey, P., & Prabhakar, R. (2016, August). An analysis of machine learning techniques (J48 & AdaBoost)-for classification. In 2016 1st India International Conference on Information Processing (IICIP) (pp. 1-6). IEEE.

Rani, K. J. (2020). Diabetes prediction using machine learning. International Journal of Scientific Research in Computer Science Engineering and Information Technology, 6, 294-305.

Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., ... & IDF Diabetes Atlas Committee. (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. Diabetes research and clinical practice, 157, 107843.

Sailasya, G., & Kumari, G. L. A. (2021). Analyzing the performance of stroke prediction using ML classification algorithms. International Journal of Advanced Computer Science and Applications, 12(6).

Sharmila, V. J., & Florinabel, J. (2021). A deep learning algorithm for COVID-19 classification using chest X-ray images. *Computational and Mathematical Methods in Medicine*, *2021*.

Sonar, P., & JayaMalini, K. (2019, March). Diabetes prediction using different machine learning approaches. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 367-371). IEEE.

Soofi, A. A., & Awan, A. (2017). Classification techniques in machine learning: applications and issues. J. Basic Appl. Sci, 13(1), 459-465.

Vijayan, V. V., & Anjali, C. (2015, December). Prediction and diagnosis of diabetes mellitus—A machine learning approach. In 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) (pp. 122-127). IEEE.

Statement and Declaration:

Conflict of interest: The author has no conflict of interest

Data Availability: The benchmark dataset is available in the Kaggle repository.

Funding Statement: The UGC-NET Junior Research Fellowship program in the Department of Computer Science, Periyar University, Salem, Tamil Nadu, India (NTA Ref. No: 230510184315, on September 05, 2023) is greatly appreciated by the first author. Ref. No: F.5-6/2018/DRS-II (SAP-II), July 26, 2018. The second author expresses gratitude to the UGC-Special Assistance Programme for funding research under the UGC-SAP at the level of DRS-II at the Department of Computer Science, Periyar University, Salem, Tamil Nadu, India.