International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 14 Issue 03, March- 2024
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 8.202 |
Thomson Reuters ID: L-5236-2015

Predicting Diabetes in Healthcare through a Hybrid Machine Learning Framework.

Name- Abhishek Johri

Guide Name - Mr.Mohd Arif

Assistant professor and Head

Department of Computer Science and Engineering

Collage Name - Rajshree Institute of Management & Technology,Bareilly

Abstract

The persistent rise in diabetes prevalence underscores the need for precise prediction models to identify individuals at risk. This research introduces a pioneering Hybrid Machine Learning Approach (HMLA) designed to elevate the accuracy of diabetes prediction within healthcare settings. Integrating diverse machine learning algorithms, including traditional statistical methods and advanced techniques like neural networks and ensemble methods, the hybrid model capitalizes on the strengths of each.dataset encompassing demographic, clinical, and lifestyle factors, this study employs feature selection techniques to optimize predictive performance and interpretability. Training the hybrid model on a sizable dataset enhances its ability to discern intricate relationships within the data.Comparative analysis demonstrates the superior predictive capabilities of the proposed approach over individual machine learning models, showcasing heightened sensitivity and specificity. Enhanced interpretability is achieved by elucidating each feature's contribution to the prediction. Rigorous validation with independent datasets affirms the robustness and generalizability of the hybrid model.This research significantly advances ongoing efforts to employ machine learning for early diabetes prediction. The presented hybrid approach stands out as a noteworthy stride in the field, offering healthcare practitioners a potent tool for prompt identification of at-risk individuals. This innovation holds promise for improving preventive healthcare strategies and, ultimately, alleviating the burden of diabetes-related complications.

Introduction

Digestion is a complex process involving the breakdown of food into essential forms of energy and nutrients. Carbohydrates, once consumed, undergo conversion into glucose within the body. Subsequently, glucose requires insulin to facilitate its journey to the cells, where it plays a crucial role in the development of tissues and muscles. The pancreas, situated behind the stomach, is responsible for insulin production.

Upon release into the bloodstream, insulin allows glucose to enter body cells. This process is pivotal as glucose serves as the primary source of energy for tissue formation and the proper functioning of body organs. However, in cases of diabetes, various factors may contribute to disruptions in this intricate system.

Insufficient insulin production by the pancreas or a lack of responsiveness from beta cells in the blood can result in elevated levels of glucose, leading to chronic diabetes issues like Type 1 and Type 2. Unlike reversible gestational and pre-diabetic conditions, Type 1 and 2 diabetes are considered irreversible. Termed the "silent killer," diabetes can give rise to various health complications, and during the pandemic, individuals with diabetes faced heightened vulnerability.Diabetic Ketoacidosis (DKA) and Hyperosmolar Hyperglycemic State are life-threatening complications of diabetes mellitus. DKA symptoms include abdominal pain, severe vomiting, excessive urination, unconsciousness, and a distinctive fruity breath odor. Treatment involves administering high volumes of fluids and insulin into the bloodstream.Timely prediction and management are crucial in mitigating the hazardous effects of diabetes. Insufficient insulin production severely impacts metabolism, and if not appropriately addressed, it triggers adverse responses in body cells. Without proper medication, vital organs such as kidneys, eyes, the cardiac system, and the nervous system can suffer substantial damage, potentially leading to organ failure and, in severe cases, death. Figure 1.1 illustrates the various categories of deaths attributable to diabetes.

Proposed Methodology

XGBoost Classifier

Boosting is a widely used technique that enhances the recognition rate of a learning algorithm by combining slightly accurate hypotheses, often referred to as weak classifiers. This approach is effective in reducing potential errors associated with weak assumptions. Adaptive boosting, developed as an iterative training method, involves delivering datasets to base learners. In this process, poorly recognized signals are assigned varying weights, denoted as 'V', with new data points being created to train subsequent weak learners in the scheme.

The weak trainer method, iterated through 'T' cycles, continually refreshes the distribution 'V' for each cycle. As the process unfolds, weights for accurately recognized sample points decrease, while those for erroneously classified data points increase. This cyclic adaptation ensures that each subsequent weak learner focuses more on misclassified points. Ultimately, the precise suggestions from these weak classifiers are combined to generate a distinctive and robust classification system.

This paper proposes the utilization of the Boost algorithm in conjunction with a weak machine learning (ML) scheme, endorsing its effectiveness in achieving optimal classification accuracy and robustness within a boosted learning framework. Adaptive boosting is advocated for integration with a supervised ML approach, maximizing efficiency through continuous monitoring, and subsequent performance evaluation. With minimal tuning constraints, Boost stands out for its speed, simplicity, and ease of programming, offering adaptability across various algorithms such as Support Vector Machines (SVM), neural networks, and random forests.

Among the most widely used algorithms for boosting in problem-solving scenarios, Gradient Boost emerges as a powerful technique that consolidates multiple "poor classifiers" into a singular "strong classifier." Its strategic emphasis on challenging-to-characterize events, while minimizing focus on well-managed ones, contributes to its high efficiency. Remarkably, Gradient Boost challenges conventional statistical norms by meticulously fitting noisy sets of data until every data point in the training set is matched without error. Even more intriguing is its continuous improvement of a strategy that already minimizes generalization errors. Adopting a statistical approach to boosting, Gradient Boost entails a step-by-step optimization of an exponential failure, involving tree regularization and step number management.

Flowchart & Algorithm

This research section introduces an Enhanced XGBoost framework, building upon the XGBoost model, often referred to as Extreme Gradient Boosting (EGB). Engineered for enhanced performance and rapidity, XGBoost is a powerful implementation of supported Decision Trees (DT). Functioning as an ensemble method and incorporating coordinated learning, XGBoost effectively combines trees to yield a more robust and well-summarized Machine Learning (ML) model.

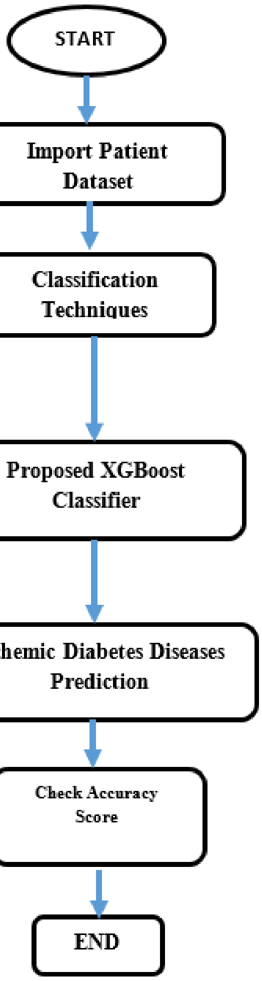

START
Import Patient Dataset
Classification Techniques
Proposed XGBoost Classifier
Ischemic Diabetes Diseases Prediction
Check Accuracy Score
END

Figure 1: Flow chart of Proposed Algorithm

The Improved XGBoost classifier integrates a variety of trees in a systematic manner to achieve superior results and accuracy. These trees are constructed sequentially, where errors are identified and rectified in each subsequent tree, leading to enhanced precision in the subsequent predictions. While the XGBoost classifier already yields favorable results in prediction, it requires more time for training in the process.the Improved XGBoost not only delivers an enhanced training model and accuracy but also addresses specific issues like tree learning. Enhanced tree learning focuses on identifying the optimal splits, and to accomplish this, a dedicated algorithm is developed, which will be detailed in the subsequent section.

Result Analysis

Machine Learning is a thought that agrees over the machine to take data from instances and former knowledge, and learn from historic data to make predictions based on the learning of the past data and that too without being programmed by any programmer i.e. we can use previous data for future predictions. In this case, instead of programmer writing the code, what a naïve user can do is feeding data to the generic algorithm, and the logic is build based on trained data by the algorithm/ machine. For e.g. When we shop online, while looking for a product, we have noticed that similar products are recommended to us to what we were looking for and we also notice the following quotation "the person who purchased this product also purchased this" type of combination of products. This recommendation is done using machine learning. Many a times we get a phone call from the bank or the finance company asking us to take a loan or purchase an insurance policy.

Figure 5.1 shows the histogram of attributes and the range of dataset attributes and code used to create it.

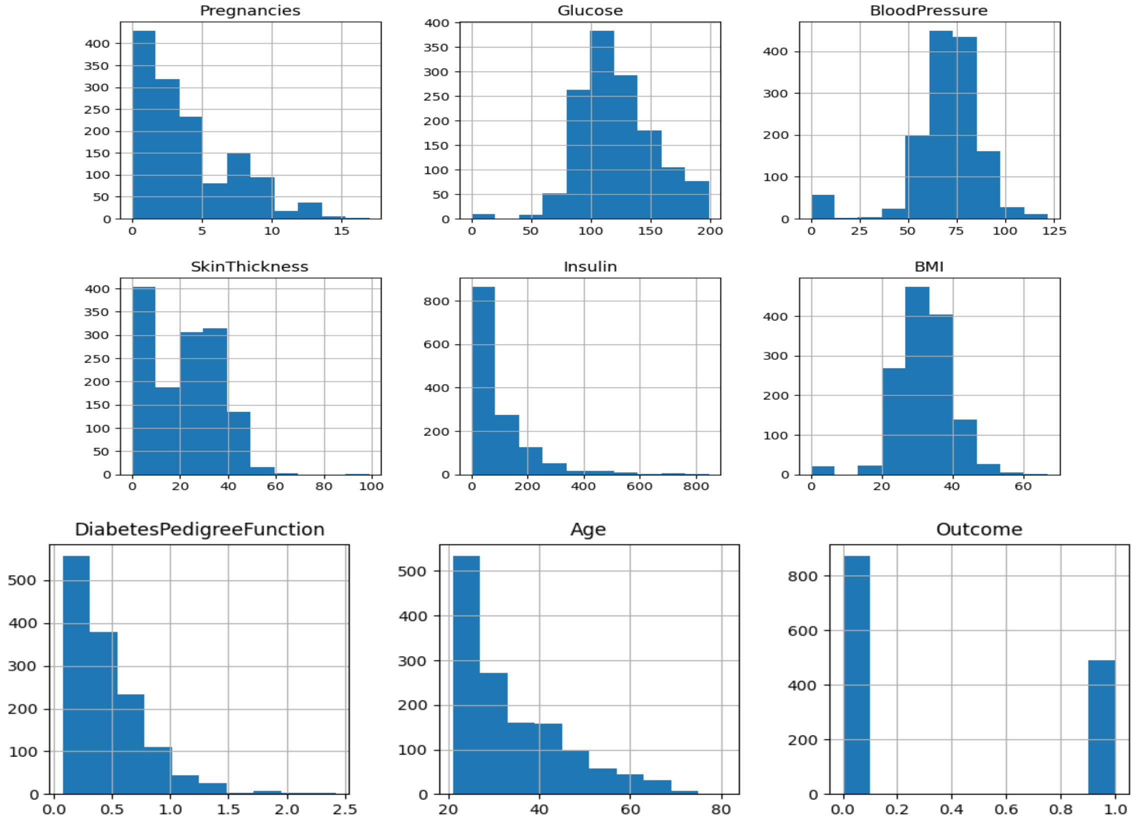


Figure 2: Histogram of Dataset

In Figures 2 and 3, the diabetes health status is visually depicted, spanning from a healthy state to severe unhealthiness. In these representations, the blue bars indicate the presence of diabetes disease, while the red bars signify the absence of diabetes disease.

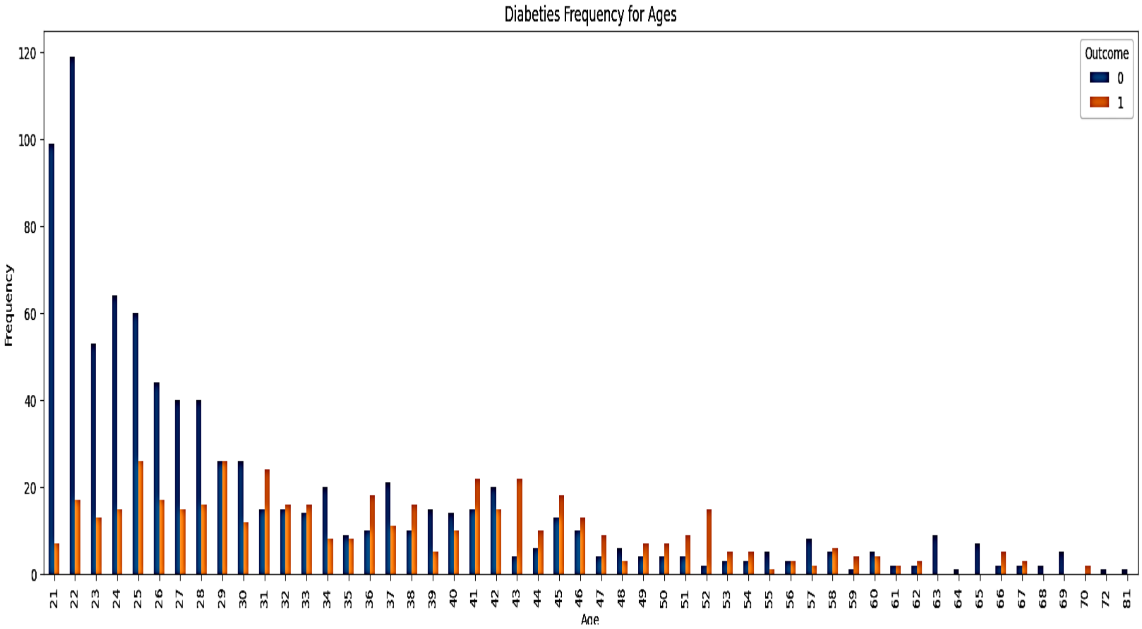


Figure 3: Bar Plot of the Number of Diabetes Frequency for Ages

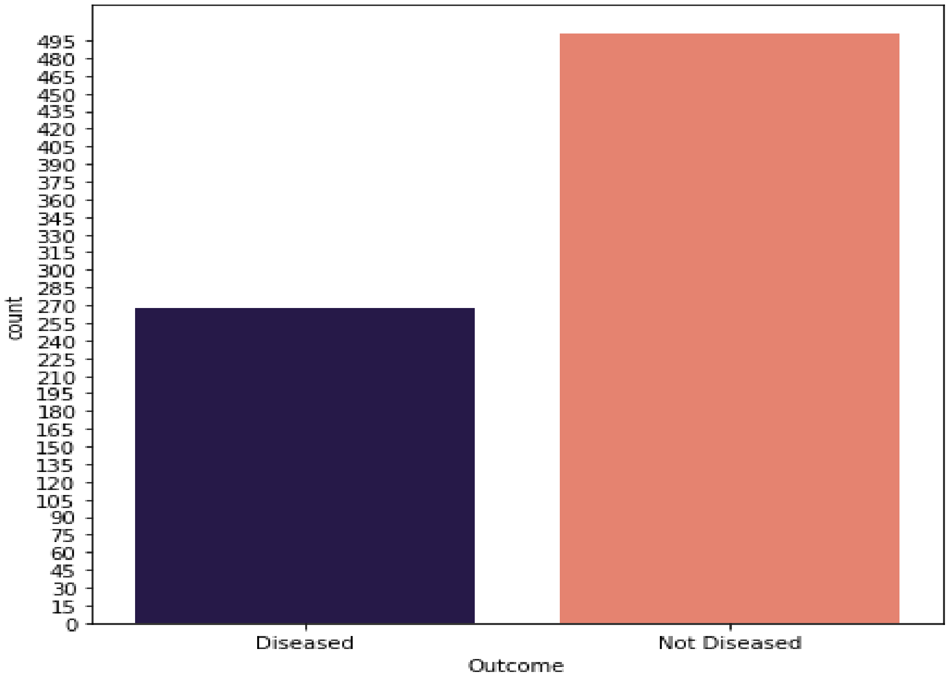


Figure 4: Bar Plot According to Outcomes

Machine learning algorithms undergo training using a designated dataset to create a model. Subsequently, when new input data with attributes is introduced to the machine learning algorithm, predictions are generated based on the established model. These predictions are then assessed for accuracy. If the accuracy of the input data meets acceptable standards, the deployment of the machine learning algorithm on the input data is executed. In cases where the accuracy of the input data falls short of acceptable levels, the algorithms undergo additional training iterations using a fresh set of data until satisfactory accuracy is achieved. This iterative training process involves refining the algorithm with new data sets to continually enhance its predictive capabilities.

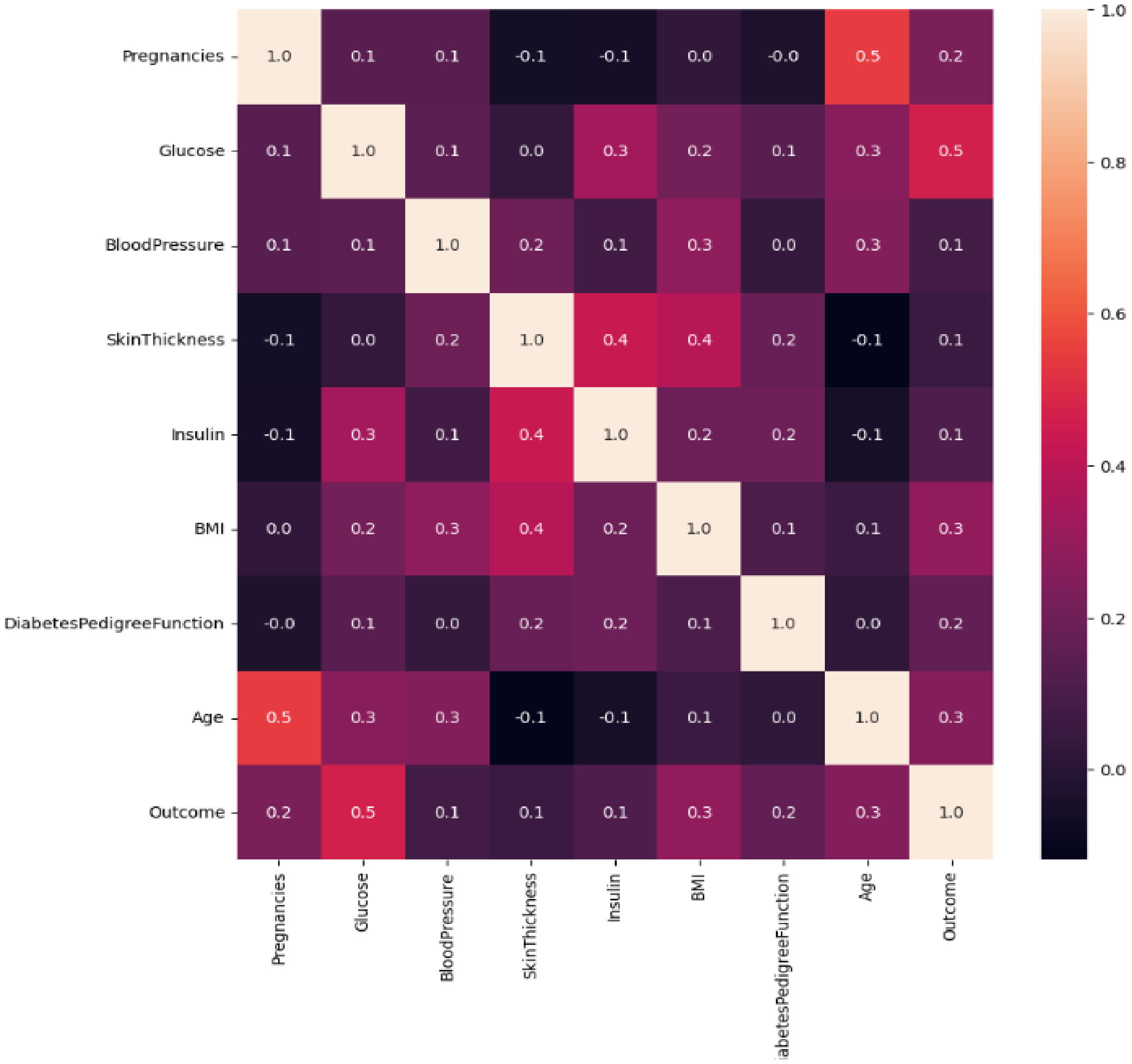


Figure 5: Bar Plot According to Diabetes Pedigree Function

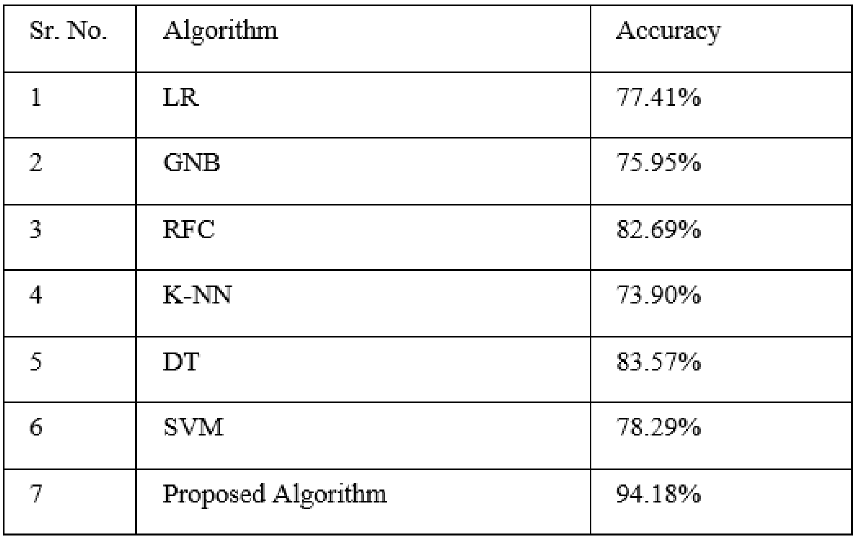
The proposed modeling approach involves the utilization of two or more distinct yet related analytical models, with their respective results combined into a unified score. In the presented algorithm, an ensemble of XGBoost classifiers was employed, achieving an impressive accuracy of 94.18%.

The Majority Vote-based model, demonstrated herein, includes classifiers such as Logistic Regression (L.R.), Naive Bayes (NB), Random Forest Classifier (R.F.C.), k-Nearest Neighbors (K.N.N.), Decision Tree (D.T.), and Support Vector Machine (SVM). This ensemble approach resulted in accuracies of 77.41%, 75.95%, 82.69%, 73.90%, 83.57%, and 78.29%, respectively, for the diabetes disease dataset.

Upon applying the machine learning approach for both testing and training, it is evident that the XGBoost classifier outperforms other methods significantly in terms of accuracy. The accuracy assessment is conducted using the confusion matrix for each algorithm, as illustrated in Figure 5. This matrix provides counts for True Positives (T.P.), True Negatives (TN), False Positives (F.P.), and False Negatives (F.N.), and the accuracy is calculated using the corresponding equation.

The results indicate that the proposed XGBoost classifier attains the highest accuracy at 94.18%, as compared to other methods. The comparative accuracy values are presented in Table 1, affirming the superior performance of the proposed approach.

Table 1: Assessment of Different Classification Methods



Sr. No.	Algorithm	Accuracy
1	LR	77.41%
2	GNB	75.95%
3	RFC	82.69%
4	K-NN	73.90%
5	DT	83.57%
6	SVM	78.29%
7	Proposed Algorithm	94.18%

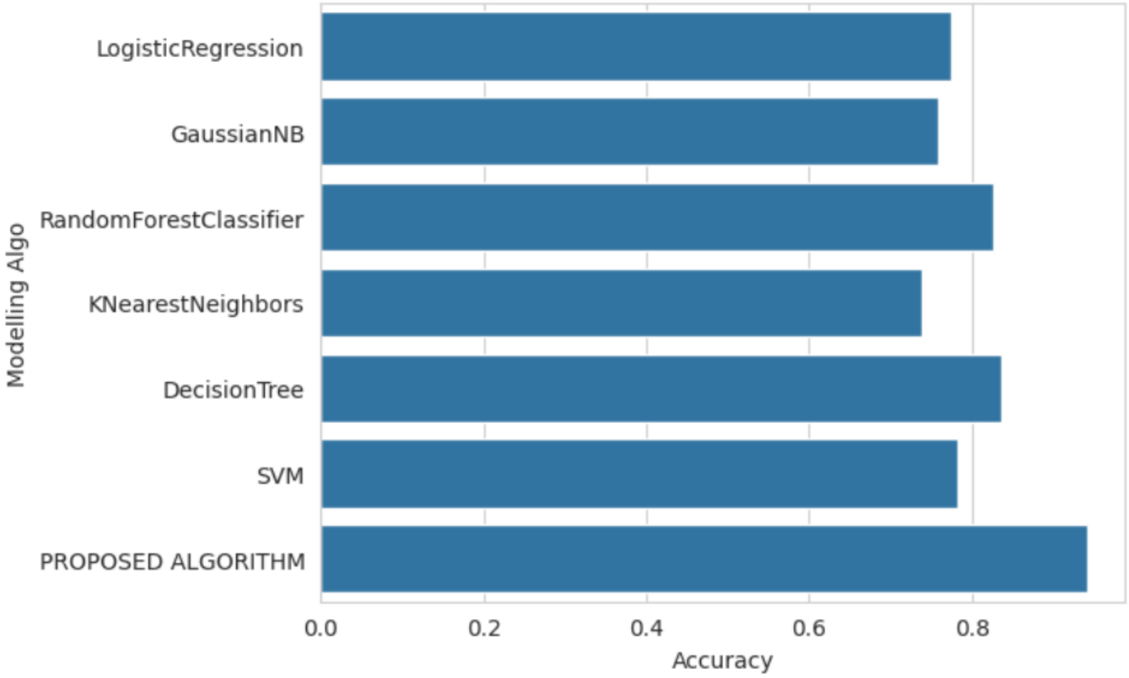


Figure 6: Bar Graph of the Different Classification Methods

Conclusion

Diabetes, a widespread metabolic disorder, necessitates robust prediction models due to its increasing prevalence and potential complications. The research proposes an innovative Hybrid Machine Learning Approach (HMLA), aiming to enhance prediction accuracy. This hybrid model strategically integrates various machine learning algorithms, including traditional statistical methods and advanced techniques like neural networks and ensemble methods. The comprehensive dataset used for training encompasses diverse demographic, clinical, and lifestyle factors, ensuring a thorough analysis.Feature selection techniques are employed to optimize the model's performance and interpretability. The hybrid model is trained on a large dataset, enabling it to capture intricate relationships within the data. Comparative analysis demonstrates the superiority of the proposed approach over individual machine learning models, showcasing heightened sensitivity and specificity. Rigorous validation with independent datasets bolsters the robustness and generalizability of the hybrid model.

References

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

[2] 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.

[3] J. B. Heaton, N. G. Polson, and J. H. Witte, "Deep learning for finance: deep portfolios," Appl. Stoch. Model. Bus. Ind., vol. 33, no. 1, pp. 3–12, Jan. 2017.

[4] Reddy S.S., Suman M., Prakash K.N. ., "Micro aneurysms detection using artificial neural networks", 2018, Lecture Notes in Electrical Engineering ,Vol: 434 ,Issue 3, pp: 409 to 417.

[5] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I., "Machine Learning and Data Mining Methods in Diabetes Research", Computational and Structural Biotechnology Journal 15, 104–116, 2017.

[6] F Mercaldo V Nardone and A Santone "Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques" Procedia Computer Science vol. 112 pp. 2519-2528 2017.

[7] I Kavakiotis O TsaveASalifoglou N Maglaveras I Vlahavas and I Chouvarda "Machine learning and data mining methods in diabetes research" Computational and structural biotechnology journal 2017.

[8] J Siryani B Tanju and TJ Eveleigh "A Machine Learning Decision-Support System Improves the Internet of Things Smart Meter Operations" IEEE Internet of Things Journal vol. 4 no. 4 pp. 1056-1066 2017.

[9] R Lafta J Zhang X Tao Y Li X Zhu Y Luo et al. "Coupling a Fast Fourier Transformation with a Machine Learning Ensemble Model to Support Recommendations for Heart Disease Patients in a Telehealth Environment" IEEE Access 2017.

[10] Majid GhonjiFeshki and OmidSojoodiShijan, "Improving the Heart Disease Diagnosis by Evolutionary Algorithm of PSO and Feed Forward Neural Network", International paper on IEEE 2016.

[11] BJ Lee and JY Kim "Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning" IEEE journal of biomedical and health informatics vol. 20 no. 1 pp. 39-46 2016.

[12] TS Brisimi CG Cassandras C Osgood IC Paschalidis and Y Zhang "Sensing and Classifying Roadway Obstacles in Smart Cities: The Street Bump System" IEEE Access vol. 4 pp. 1301-1312 2016.

[13] L. Hermawanti, "Combining of Backward Elimination and Naive Bayes Algorithm To Diagnose Breast Cancer", Momentum, vol. 11, no. 1, pp. 42-45, 2015.