Advances in Parallel Computing Algorithms, Tools and Paradigms
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC220021

Comparative Analysis of Predicting the Diabetic Disease Using Machine Learning Techniques

J Revathy[a,1] and D Selvanayagi[b]
[a]*Research scholar, Department of Computer Applications*
[b]*Assit. Prof, Department of Computer Applications*
[a.b]*Vellalar College for women, Erode, Tamil Nadu, India*

Abstract. Machine Learning is concerned with the making of calculations and methods that use PCs to learn and acquire insight, using the related knowledgeavailable.This work is focused on machine learning approaches for predicting diabetic disorders, using datasets from Predict the Diabetic Diseases. A web-based comparative analysis of multiple machine learning algorithms (Decision Tree, Support Vector Machine, K- Nearest Neighbor, and Logistic Regression) is utilized in this paper, to assess their performances in recognizing reliable models for detecting diabetic disease. To see the effects of adding more features to the classification model, three performance measures were chosen: F1-Measure, Precision, and Accuracy.

Keywords. Machine Learning, Decision Tree, Support Vector Machine, Logistic Regression, Diabetes

1. Introduction

Diabetes diagnosis has been carried out utilizing multiple machine learning algorithms to forecast the illness of a patient and to improve treatment such as Decision Tree (DT), Support Vector Machine (SVM), *k*- Nearest Neighbor (*k*-NN), and Logistic Regression. Diabetes Mellitus (DM) is an ongoing condition described by the powerlessness of the body to use glucose. Early discovery of the illness brings down clinical usage and the probability of individual's genuine medical issues. DM research is a critical methodology for separating the needed information from the gigantic database of diabetes-related information. The financial cost is one of the main concerns in clinical science study, and it normally creates enormous volumes of information.

Types of Diabetes

Type 1 Diabetes - The body doesn't fabricate sufficient insulin in this kind of diabetes. Insulin-subordinate diabetes, adolescent diabetes, and beginning stage diabetes are the terms used to depict this sort of diabetes. Type 1 diabetes, for the mostly part,

[1] J Revathy, Department of Computer Applications, Vellalar College for women, Erode, Tamil Nadu, India,; E-mail: revamsc12@gmail.com.

occursbefore the age of 40, for example in early adulthood or during puberty, and the patient requires insulin shots for life.
Type 2 Diabetes - The body doesn't make sufficient insulin or the cells in the body are insulin safe. A few people with type 2 might have the option to control their indications by reducing body weight, following a decent eating regimen, and intently checking their blood glucose levels. AsType 2 diabetes deteriorates after some time, the patient will very likely need to take insulin, for the most part in tablet structure. Obesity, absence of physical work, and eating unhealthy food lead to type 2 diabetes.

Type 3 Diabetes (Gestational Diabetes) –Women develop this during pregnancy. A few ladies' blood glucose levels rise since their body doesn’t produce sufficient insulin to move all of the glucose into their cells, bringing about consistently rising glucose levels. Most of ladies, with gestational diabetes, might deal with their diabetes with physical work and a sound eating routine. About 10% and 20% of them may require blood glucose controlling medicine.

The focus of this study is to review various diabetes detection mechanisms on the Type 3 diabetes to identify this, Naive Bayes (NB), SVM, and DT classification methods are employed and assessed on Pima Indian Diabetic dataset (PIDD). Using several criteria, the experimental performances of these algorithms are compared, and they all attain good accuracy [1].This study examines the familiar metrics such as F1-score, Precision, Recall and the Accuracy rate of various diabetes detection mechanisms [2].

2. Related Work

The research work [3] has collected the diabetic dataset, and various machine learning techniques are implemented. The Logistic Regression provides the highest accuracy of 96%and AdaBoost is found to be the best model with an accuracy rate of 98.8%. With two separate datasets, the accuracies of machine learning algorithms are compared.

Calisiretal [4] surveyed diabetic illnesses and diagnosed them using the Linear Discriminant Analysis (LDA) and Morlet Wavelet Support Vector Machine (MWSVM) classifiers. The LDA – MWSVM automatic diagnosis method for diabetes achieves extremely promising results in classifying possible diabetic individuals. The employed LDA – MWSVM automatic diagnosis system is thought to be particularly useful to clinicians, in making final decisions about their patients.

Georga et al [5] applied its capability to the table for smooth, worldwide, and meager answers for nonlinear regression problems. The Support Vector Regression (SVR)approach is pushed for multivariate prescient investigation of glucose digestion. They estimated the impact of each contribution on forecast exactness, utilizing different information circumstances, and utilizing the present moment and the long haul expectations. Finally, in addition to predicting values, the challenges of anticipating hypo/hyperglycemia episodes are also warned during experimentation.

Geetha. G et al [6] have utilized Random Forest (RF)for the prediction of diabetics. After separating and analyzing the training and testing data, this method provides us an approximate result. When compared to Nave Bayes, this approach has a substantially higher efficiency. As a result, the suggested method is highly effective for both diabetic analysis and prediction.

Deberneh et al [7]had indicated Type 2 diabetes occurrence prediction model that predicts as normal, pre-diabetes, or diabetes in the following year (Year+1). The forecast models were made utilizing LR, RF, XGBoost, SVM, and group classifiers (CIM, ST, and SV). FPG, HbA1c, fatty substances, BMI, gamma-GTP, orientation, age, uric corrosive, smoking, drinking, actual work, and family ancestry were among the elements picked. The model gives critical data on the event of type 2 to the two specialists and patients early.

Kavakiotis et al [8] have surveyed many machine learning and data mining approaches used in DM. This work includes clinical and biological information that leads to more in-depth exploration of diagnosis. Polat et al [9] have depicted the impacts of master frameworks and AI methods for the analysis of diabetes. Settling on choices in the clinical business can be troublesome at times now and then. Clinical dynamic order frameworks permit clinical information to be investigated significantly quicker and in more detail. An interesting cross breed approach in the light of PCA and ANFIS is proposed in this paper for the determination of diabetes disease. The goal of diagnosing diabetics was applied to ANFIS with PCA in the exploration revealed in this work, and the most reliable learning approaches were assessed. On the diabetes sickness dataset, tests were embraced to analyze diabetes illness naturally, utilizing ANFIS and PCA.

Alghamdi et al [10] have applied AI techniques for anticipating occurrence diabetes, utilizing cardiopulmonary wellness information. Among the five characterization models utilized, the Random Under-Sampling procedure yielded improvement. The SMOTE approach, then again, exhibited an impressive improvement in the forecast capacity of all characterization models, as the rates used were step by step expanded. In every one of the models, the Random Forest and NB Tree models beat the others (Kappa, Recall, Precision and Specificity).

Faruque et al [11] have examined the early discovery of diabetes by thinking about a few gamble factors related with the disease. Foreseeing diabetic onset can be pretty much as straightforward as removing information from a genuine medical services dataset. To estimate diabetes mellitus, tests were directed on grown-up populace, utilizing four normal AI calculations SVM, NB, k-NN, and C4.5. The exhibition of the C4.5 is a lot better than the other AI calculations, for the arrangement of diabetic information. Soni et al [12]have applied machine learning algorithm to design and implement diabetes prediction such as SVM, *k*-NN, RF, DT, LR, and Gradient Boosting classifiers. The exploratory outcomes can assist the patient'swell-being, with caring experts settling on early forecasts and choices, to fix diabetes and save individuals lives.

Kadhm et al [13] have proposed a fast and accurate diabetes prediction algorithm. For each PID dataset, the suggested system used 768 instances with 8 attributes. Furthermore, by partitioning the dataset into subsets, an optimal classification result was achieved. The suggested approach concentrated on the analysis and classification of characteristics. The results, of the studies, demonstrated the benefits of using the suggested system's algorithms, by achieving a greater categorization rate than competing systems.

Tigga et al [14] have applied six different AI, and the outcomes were thought about. Tests were run on information gained on the web, and disconnected surveys, that included 18 diabetes-related inquiries. The exactness of RF is 94.10%, which is the best among the others. For the PIMA dataset, RF likewise, gives the greatest exactness. Each of the models showed great outcomes for greater accuracy, review responsiveness, etc.

Nnamokoetal [15] have collected a variety of factors that influence their potential to improve performance, particularly at the base level. The basis classifiers are chosen from five different machine learning algorithm families. As a result, each classifier would generate models based on its operational features. Despite the fact that none of them produced any major improvements at the base level, the sum of their unique biases contributed to a greater understanding of the classification problem at the ensemble level, ultimately leading to significant improvement. If such pre-processing experiments had been carried out, their outcomes might have been different. However, the findings of this study show that advances were achieved as a result of feature selection applied to heterogeneous base learners, rather than data pre-processing.

Rajeswari et.al [16] has splitted the mathematical properties (hazard factors) of the PID informational collection, the recommended strategy - Fuzzy rationale based Associative Classification to resolve the issue of limit esteem confusion. It doesn't require proficient parceling or reach fixing information. This is reachable in the light of the fact that the MFs utilize normal limit parting to break the base and the most extreme worth of the characteristics into different language phrases. Whenever the idea of the FARS, created by the recommended strategy is contrasted with the exception pruning measurements - Lift, Leverage, and P-Value, obviously the proposed technique delivers a decent model. The proposed procedure is superior to the fresh strategy in deciding the specific gamble factors like Age, Glucose, DPF, BMI, and BP, as well as their perilous qualities, to foresee prediabetes.

Agnal and Saraswathi[17] have applied Naive Bayes Classifier's methodological approach to reliably and precisely diagnose diabetes. The proposed approach outperforms Hadoop Technology, in handling both large and small volumes of data. As a classification algorithm, this procedure is quick, secure and easy to change, and produces accurate results. By improving accuracy, it can be beneficial in medical research.

VijiyaKumar et al [18] have introduced determination of a hereditary anomaly at the beginning phase, as one of the necessary genuine clinical worries. Irregular Forecast calculations have been analyzed and evaluated on an assortment of scales during this undertaking. The objective of this venture is to make a framework that can precisely gauge diabetes in a patient early, using an AI strategy that gives advance assistance to anticipating diabetes exactness rates. The consequences of different AI calculations are displayed in Table 1, for the PIMA Indian Diabetic Datasets.

Table 1. Performance of various Machine Learning algorithms on PIMA Dataset

Algorithm	F1-Score	Precision	Accuracy	Recall
Linear regression	0.722732	0.754323	0.843782	0.693683
Logistic Regression	2.313075	0. 771025	0.884578	0.761037
Decision Tree	2.648536	0.672584	0.703565	0.662134
SVM	2.627466	0.875822	0.917324	0.846212
Naive Bayes	3.164824	0.800103	0 .815395	0.791206
k-NN	1.129604	0.274685	0.318189	0.282401
Random Forest	1.075652	0.265264	0.282280	0.268913

3. Conclusion

The first point of this advancement was to propose and achieve Diabetes Prediction Using Machine learning strategies and investigation of a strategy has been classified. The proposed methodology approach uses different categorization such as Linearegression, Decision Tree, Logistic Regression, SVM, Naive Bayes, KNN, random Forest classifiers are used. The comparative analysis and results illustrate the early way to cure diabetes on earlier prediction to save a human life with machine learning Techniques.

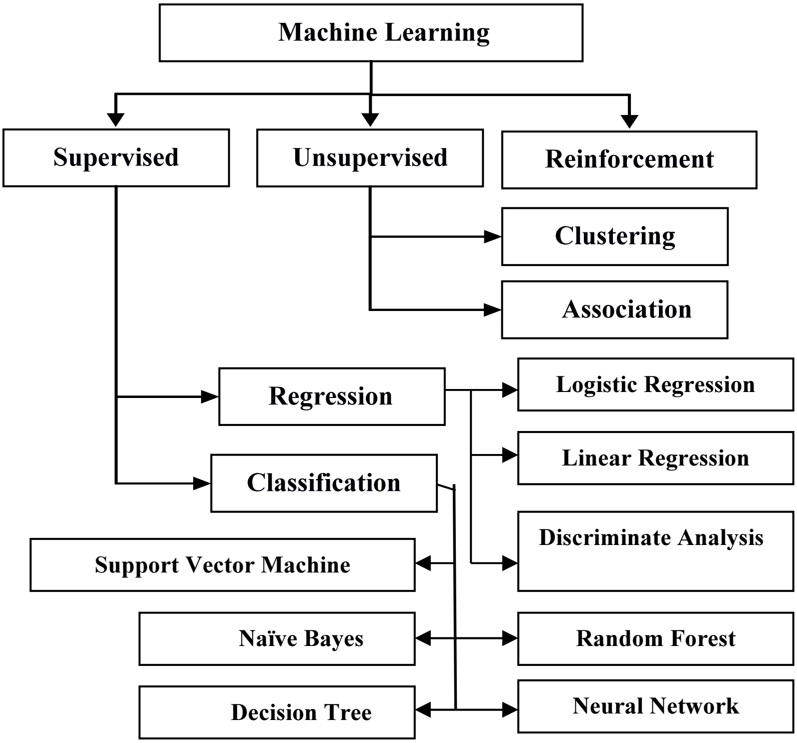

Figure 1. Various Machine Learning Algorithm used for diabetes prediction.

References

[1] Sisodia, D., &Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia computer science, 132, 1578-1585.

[2] Anusha, B., S. Sabena, and L. Sairamesh. "Optimized Food Recognition System for Diabetic Patients."In International Conference on Next Generation Computing Technologies, pp. 504-525. Springer, Singapore, 2017.

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

[4] Çalişir, D., &Doğantekin, E. (2011). An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier. Expert Systems with Applications, 38(7), 8311-8315.

[5] Georga, E. I., Protopappas, V. C., Ardigo, D., Marina, M., Zavaroni, I., Polyzos, D., & Fotiadis, D. I. (2012). Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression. IEEE journal of biomedical and health informatics, 17(1), 71-81.

[6] Geetha, G., & Prasad, K. M. (2020). Prediction of Diabetics using Machine Learning. IJRTE (International Journal of Recent Technology and Engineering), 8(5), 1119-1124.

[7] Deberneh, H. M., & Kim, I. (2021). Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. International Journal of Environmental Research and Public Health, 18(6), 3317.

[8] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., &Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15, 104-116.

[9] Polat, K., &Güneş, S. (2007). An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. Digital Signal Processing, 17(4), 702-710.

[10] Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., &Sakr, S. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. PloS one, 12(7), e0179805.

[11] Faruque, M. F., &Sarker, I. H. (2019, February). Performance analysis of machine learning techniques to predict diabetes mellitus. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-4). IEEE.

[12] Soni, M., & Varma, S. Diabetes Prediction using Machine Learning Techniques.International Journal of Engineering Research & Technology (IJERT), 9 (9), September 2020. (11)

[13] Kadhm, M. S., Ghindawi, I. W., &Mhawi, D. E. (2018). An accurate diabetes prediction system based on K-means clustering and proposed classification approach. International Journal of Applied Engineering Research, 13(6), 4038-4041.

[14] Tigga, N. P., & Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science, 167, 706-716.

[15] Nnamoko, N., Hussain, A., & England, D. (2018, July). Predicting Diabetes Onset: An Ensemble Supervised Learning Approach. In 2018 IEEE Congress on Evolutionary Computation (CEC) (pp. 1-7). IEEE.

[16] Rajeswari, A. M., Sidhika, M. S., Kalaivani, M., &Deisy, C. (2018, April). Prediction of prediabetes using fuzzy logic based association classification. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 782-787). IEEE.

[17] Agnal, A. S., &Saraswathi, E. (2020). Analyzing Diabetic Data Using Naive-Bayes Classifier. European Journal of Molecular & Clinical Medicine, 7(4), 2687-2699.

[18] VijiyaKumar, K., Lavanya, B., Nirmala, I., & Caroline, S. S. (2019, March). Random forest algorithm for the prediction of diabetes. In 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN) (pp. 1-5). IEEE.