IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 10, Issue.5, May 2022 E-ISSN: 2347-2693

Python Based Diabetes Prediction Using Ensemble Machine Learning Techniques Using LR Algorithm and Hybrid Method

Pradeep Kumar G.[1*], R. Vadivel[2]

[1]PG Student Department of Information Technology, Bharathiar University, Tamil Nadu India
[2]Assistant Professor Department of Information Technology, Bharathiar University, Tamil Nadu, India

**Corresponding Author: pradeepganeshanpoy@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v10i5.4346 | Available online at: www.ijcseonline.org

Received: 21/Apr/2022, Accepted: 08/May/2022, Published: 31/May/2022

***Abstract*—** The constant flood of fresh patient data is causing problems in the healthcare system. Researchers have been utilizing this data to help the healthcare industry improve its capacity to manage major diseases. They are also looking at how patients might be informed of symptoms in a timely way, therefore avoiding the serious hazards that come with them. Diabetes is one such condition that is spreading at an alarming rate these days. It may lead to a number of significant problems, such as decreased eyesight, myopia, burning extremities, renal failure, and heart failure. When blood sugar levels rise over a certain threshold, the human body is unable to manufacture enough insulin to maintain the appropriate level. As a consequence, diabetics must be educated on the need of adhering to appropriate treatment regimens. As a consequence, early diabetes diagnosis and classification are crucial. This method employs Machine Learning approaches to improve diabetes prediction accuracy. Furthermore, the trials showed that ensemble classifier models outperformed base classifier models on their own. Its results were compared to the same dataset when various classification techniques such as random forest, support vector machine, decision tree, and naive bayes were applied to it.

***Keywords*—**ML, Diabetic Prediction, SVM, DT, ND, LR, Ensemble

I. INTRODUCTION

Numerous opportunities in healthcare have arisen as a consequence of the better predictive analytics afforded by machine learning models. There are machine learning algorithms that can forecast chronic illnesses including heart disease, infection, and intestinal sickness. Furthermore, there are several new machine learning models for forecasting noncommunicable diseases that will continue to provide value to healthcare. Researchers are working on machine learning models that will offer an incredibly early prediction of a specific ailment in a patient, allowing for the creation of effective disease preventive strategies. This will also result in fewer hospitalizations for patients. This change will have a huge beneficial influence on healthcare organisations.

The area of healthcare research that draws the most interest is the healthcare system that makes use of modern computer technologies. As previously stated, relevant academics are already partnering with healthcare companies to develop more technologically advanced solutions. Diabetes is a disease in which the body's ability to produce insulin is compromised. In other words, the body is unable to respond to the production of the hormone insulin. As a result, carbohydrate metabolism is disrupted, and blood glucose levels rise. Diabetes detection at an early stage is crucial for the reasons indicated before. Diabetes affects a vast number of people globally, and the number is increasing on a regular basis. Because this disorder has the potential to affect multiple vital organs, early detection will help the medical community address it. As the number of diabetic patients grows, it is necessary to save vital medical data. Researchers must build a system for collecting, managing, and analysing diabetic data, as well as spotting possible dangers, with the help of evolving technologies.

Diabetes causes blood glucose levels to get unusually high. The body produces glucose as a function of food ingestion. Insulin is a hormone produced by the body that aids in the balance of glucose levels and the regulation of blood sugar levels; insulin deficiency leads to diabetes. Type 1 diabetes is a disorder in which the body does not produce enough insulin to keep blood sugar levels normal. Type 2 diabetes develops when the body produces insulin but does not completely use it in order to maintain appropriate blood sugar levels. Kind 2 diabetes is the most common type. Prediabetes is a condition in which a person's glucose level is high but not high enough to be diagnosed with diabetes. Individuals with prediabetes, on the other hand, are susceptible to developing type 2 diabetes. This illness has the potential to wreak havoc on a number of vital organs, including the kidneys, heart, nerves, and eyes. Gestational diabetes happens when a woman acquires the disease while pregnant. Diabetes may be controlled if we maintain a healthy weight, consume a well-balanced diet, and exercise frequently. Keep an eye on one's blood sugar levels at all times.

II. LITERATURE SURVEY

This chapter discusses the research on the detection and diagnosis of diabetes mellitus that has been undertaken

utilising different approaches and methodologies outlined in the published literature. The current approaches and processes for diagnosing diabetes mellitus, as well as their limitations, are described briefly here.

Patil et al. (2010) developed a novel technique for generating association rules for the categorization of Type-II diabetes. On the basis of the numeric data, it constructed association rules. Initially, this technique began with pre-processing the data, which addresses the Dataset's missing values. Then, a modified equal width bin interval was developed, which discreteness the Dataset's continuous-valued properties. The estimated width of the required gap was determined and supplied to the model based on the judgment of medical professionals. The numeric characteristics in the Dataset were transformed to categorical values using the modified equal width binning interval in this model. After transforming the continuous variable to categorical values, the Apriori technique was used to determine the variables' hidden link. The Apriori algorithm-generated criteria defined and incorporated all possible combinations of risk variables that resulted in diabetes or did not result in diabetes within five years. Association rules accurately predicted the diabetes dataset's class labels. However, the computational cost of generating the association rule is considerable.

A Reinforcement Learning-based Evolutionary Fuzzy Rule-Based System (RLEFRBS) was created for the early detection of diabetes (Mansourypoor & Asadi 2017). The RLEFRBS methodology included developing a Rule Base (R.B.) and optimising rules. Without starting rules, an initial R.B. was generated using the Dataset's numerical data. The redundant rules were then deleted depending on the confidence level. By removing superfluous criteria in the preceding sections, clearer rules with increased interpretability were obtained. Finally, a genetic algorithm is used to pick subsets of rules that are more useful to categorizing the data. The specified system's performance was enhanced by evolutionary tuning of the membership functions and weight adjustment using reinforcement learning. The suggested approach is intended to be used just for the diagnosis of type-2 diabetic illness.

Hayashi and Yukita (2016) developed a Sampling Recursive Rule Extraction (Re-RX) method with the J48graft algorithm for diagnosing type 2 diabetes mellitus in the Pima Indian diabetes dataset. The recursive rule extraction model was a white model that delivered a high degree of classification accuracy. This method produced more rules than the previous algorithm due to its recursive nature. To generate a classification rule that is very accurate, simple, and interpretable, the rule extraction method Re-RX with J48graft was suggested. It used a combination of sample selection approaches. It is proved that rules generated from the Pima Indian Diabetes Database using the two types of tradeoffs were more accurate, succinct, and interpretable, and hence more acceptable for medical decision making.

For diabetes diagnosis, a hybrid feature selection strategy based on the weighted least squares twin support vector machine approach (WLSTSVM) was presented (Tomar & Agarwal 2015). The suggested model used the WLSTSVM as an examination technique; sequential forward selection (SFS) was used to conduct the search; and correlation feature selection (CFS) was used to determine the relevance of the highlights. This approach efficiently identified key characteristics and circumvented the issue of class imbalance. This model selected more relevant characteristics from the Dataset using hybrid feature selection. The hybrid feature selection approach was an effective strategy for selecting features that incorporated the advantages of filter-based and wrapper-based feature selection. The performance of the HFS-based WLSTSVM approach was evaluated using predicted accuracy, sensitivity, specificity, and geometric mean on three well-known illness datasets obtained from the UCI repository.

III. DESIGN AND DEVELOPMENT

Technology advances at a breakneck pace, making it possible to apply machine learning to data and anticipate the outcomes, which becomes a classification challenge in machine learning. For example, there are two forms of diabetes; cancer has four stages; and so on. Diabetes is a chronic condition in which the body's insulin levels decline and glucose levels exceed the threshold limits. There are two forms of diabetes. Type 1 diabetes affects 5-10% of people because their cells do not react to insulin, whereas Type 2 diabetes affects 90-95 percent of people because their cells do not respond to insulin levels. If not treated promptly, the condition might progress to further cardiovascular and renal issues. The present model's disadvantage is the pre-processing approach, since the model has a 268:500 ratio of diabetic to non-diabetic data.
This study is based on the prediction of diabetes. The diabetes prediction approach consists of three stages: pre-processing, determining the significance of features, and classification. The suggested technique's steps are detailed below:-

A. Dataset Description-

The data is gathered from the Kaggle website https://www.kaggle.com/uciml/pima-indians-diabetes-database/version/1 which is named as Pima Indian Diabetes Dataset. The Dataset has many attributes of 768 patients.

B. Data Preprocessing-

The most critical procedure is data pre-processing. The majority of healthcare data has missing values and other contaminants that might impair the data's efficacy. Pre-processing data enhances the quality and efficacy of the results acquired throughout the mining process. Effectively applying Machine Learning Techniques to the Dataset is critical for accurate findings and good prediction. Two steps are required to do pre-processing on the Pima Indian diabetes dataset.

Remove all occurrences having a value of zero (0). It is not possible to have a value of zero. As a result, this occurrence is omitted. By removing unimportant features/instances, we create a feature subset, which decreases the dimensionality of the data and enables quicker processing.

Splitting data- Once the data has been cleaned, it is standardized for training and testing the model. When data is divided, the algorithm is trained on the training set while the test set is kept separate. This training method generates the training model based on the logic and algorithms of the features, as well as the values in the training data. Normalization seeks to harmonies all of the qualities.

C. Feature Importance

The second step will implement the method of feature significance in the network. The feature importance method establishes the link between the attribute and the target set. The R.F. method is applied to the input dataset, which we will refer to as the Dataset. The L.R. method will forecast the most often occurring characteristic in the Dataset.

After pre-processing the dataset the features are selected by using hybrid LR and
RF algorithm with the 1000 samples and 10 selected features.

D. Ensembling

Ensemble is a Machine Learning approach that use meta-algorithms to integrate many machine learning techniques into a single optimum predictive model in order to minimize variance, bias, or enhance predictions. When compared to a single model, this technique results in enhanced predictive performance. Numerous assembly techniques exist, including bagging, boosting, ada-boosting, stacking, voting, and averaging. We used a voting-based assembly strategy to assemble the PIMA Indian diabetes dataset.

SVM, DT, RF, KNN, and NB classification algorithms combine and use stacking ensemble classifier algorithms. All of the code was written in Python. Python libraries are used in our source code: Keras, NLTK, Numpy, Pandas, Sklearn, and scikit. Algorithms were judged on their accuracy, precision, recall, and F-score, among other metrics.

After the feature selection the classification has done with the use of ML algorithms in separate manner. Then the stacking ensemble method for classification process to classify the diabetes details. Like precision, recall, f-measure and support.

IV. RESULT AND DISCUSSION

The Implementation has successfully carried out in python programming language. The various ML algorithms has implemented and tested. Like DT, SVM, LR, NB and Hybrid Method.

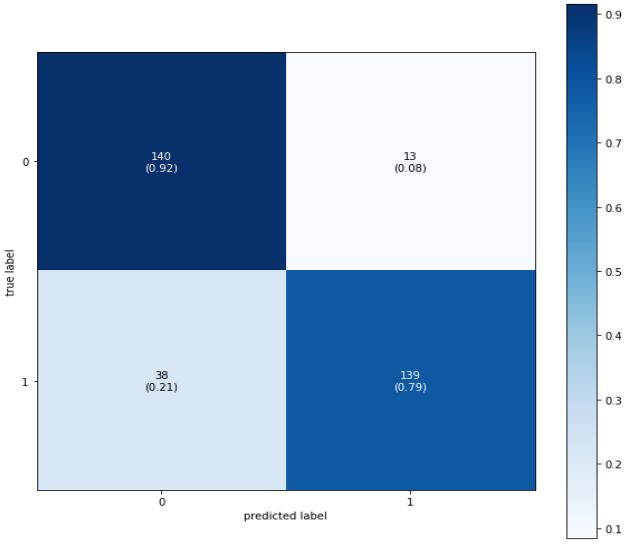

Figure 1: Hybrid Confusion matrix

Figure 1: illustrates the Hybrid confusion matrix for the stacking ensemble algorithm used, which embeds by each algorithm. The TP is achieved 140 and TN is 139. And FP is 13, FN is 38.

Table 1: Comparison table for various ML model with Proposed DPE

Model	Accuracy (%)	Precision	Recall	F-measure
SVM	93	77	91	83
DT	86	88	83	85
LR	84	79	92	85
NB	77	73	87	77
DPE (Hybrid)	95.5	85	89	87

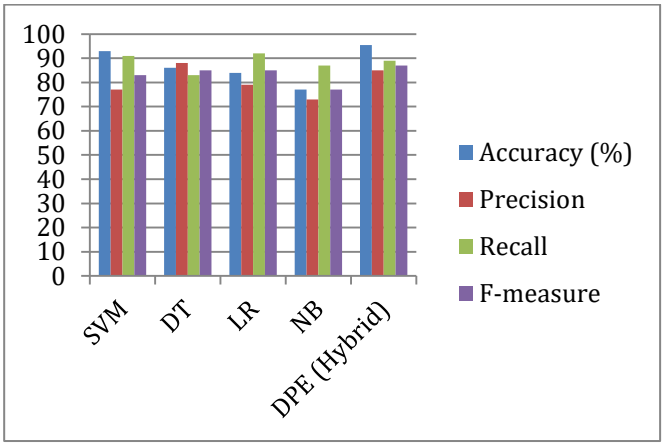

Figure 2: Performance metrics as accuracy, precision, recall, F-measure

The table 1 shows the performance metrics about various algorithms. And figure 2 respond to the comparison chart. While SVM is achieved accuracy, precision, recall, f-

measure 93, 77, 91, 83 respectively. In DT using 86, 88, 83, 85.
In LR used 84, 79, 92, 85. In NB used 77, 73, 87 and 77. Finally the DPE method has used 95.5, 85, 89, 87 achieved as accuracy, precision, recall, f-measure values.

V. CONCLUSION

The goal of this study is to develop a machine learning-based technique for predicting early diabetes using computer-aided diagnostic tools. Numerous solutions have been developed in this field, however attaining accuracy and computing complexity remains challenging. As a consequence, an ensemble classifier model is created in which feature selection, feature ranking, and a neural network-based approach are integrated to learn feature and category patterns efficiently. The experimental results show that the proposed technique enhances classification accuracy. In the future, if we have a large dataset of diabetic patients, we may do a comparative study to assess the performance of each algorithm and the hybrid algorithm, with the objective of deciding which algorithm is superior for predictive analysis. A particular technique for identifying diabetes is not a complex strategy for early diagnosis, and it is not totally dependable for disease prediction. We want an advanced hybrid predictive analytics diabetes diagnostic system that is both accurate and efficient. We can utilise data mining, which is a neural network for studying and exploiting, to help doctors make medical decisions that raise the chance of pregnant diabetes diagnosis. We are unable to predict the type of diabetes since the datasets we presently have are not accurate. In the future, we want to forecast and explore the type of diabetes, which may improve diabetes prediction accuracy. In addition, we may look at the causes of diabetes and ways to avoid getting it.

REFERENCES

[1] Adarsh, P & Jeyakumari, "Multiclass SVM-based automated diagnosis of diabetic retinopathy", Communications and Signal Processing (ICCSP), International Conference on IEEE, **pp. 206-210, D 2013.**

[2] Akram, MU, Tariq, A, Khan, SA & Bazar, "Microaneurysm detection for early diagnosis of diabetic retinopathy", Electronics, Computer and Computation (ICECCO), International Conference on IEEE, **pp. 21-24, SA 2013.**

[3] Amin, J, Sharif, M, Yasmin, M, Ali, H & Fernandes, "A method for the detection and classification of diabetic retinopathy using structural predictors of bright lesions", Journal of Computational Science, **vol. 19, pp. 153-164, SL 2017.**

[4] Patil, BM, Joshi, RC & Toshniwal, "Association rule for classification of type-2 diabetic patients", Machine Learning and Computing (ICMLC),Second International Conference on IEEE, **pp. 330-334, , D 2010.**

[5] Mansourypoor, F & Asadi, "Development of a reinforcement learning-based evolutionary fuzzy rule-based system for diabetes diagnoses", Computers in Biology and Medicine, **vol. 91, pp. 337-352, S 2017.**

[6] Hayashi, Y & Yukita, "Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset", Informatics in Medicine Unlocked, **vol. 2, pp. 92-104, S 2016.**

[7] Tomar, D, & Agarwal, "Hybrid feature selection based weighted least squares twin support vector machine approach for diagnosing breast cancer, hepatitis, and diabetes", Advances in Artificial Neural Systems, **vol. 2015, pp. 1-10, S 2015.**

[8] Lukmanto, RB, Nugroho, A & Akbar, "Early detection of diabetes mellitus using feature selection and fuzzy support vector machine", Procedia Computer Science, **vol. 157, pp. 46-54, , H 2019.**

[9] Vijayan, VV & Anjali, "Prediction and diagnosis of diabetes mellitus - A machine learning approach", Intelligent Computational Systems (RAICS), 2015 IEEE Recent Advances in IEEE, **pp. 122-127, C 2015.**

[10] Yıldırım, EG, Karahoca, A & Uçar, "Dosage planning for diabetes patients using data mining methods", Procedia Computer Science, **vol. 3, pp. 1374-1380, T 2011.**

[11] Zarkogianni, K, Litsa, E, Mitsis, K, Wu, PY Kaddi, CD, Cheng, CW & Nikita, "A review of emerging technologies for the management of diabetes mellitus", IEEE Transactions on Biomedical Engineering, **vol. 62, no. 12, pp. 2735-2749, KS 2015.**

[12] Zhang, B Kumar, BV & Zhang, "Detecting diabetes mellitus and non proliferative diabetic retinopathy using tongue color, texture, and geometry features", IEEE Transactions on Biomedical Engineering, **vol. 61, no. 2, pp. 491-501, D 2014.**

[13] Yookesh, T. L., et al. "Efficiency of iterative filtering method for solving Volterra fuzzy integral equations with a delay and material investigation." Materials today: Proceedings 47: 6101-6104, **2021.**

[14] Kumar, E. Boopathi, and V. Thiagarasu. "Segmentation using Fuzzy Membership Functions: An Approach." IJCSE, ISSN: 2347-2693, **2017.**

AUTHORS PROFILE

Mr. Pradeep Kumar.G received Bachelor's Degree in Computer technology in the year 2020 from Sri Krishna Arts and Science college, Coimbatore, Tamil Nadu, affiliated to Bharathiar University. He is currently pursuing a Master's Degree in Information Technology from 2020 to 2022, at Bharathiar University, Coimbatore, Tamil Nadu.

Dr. R.Vadivel is an Assistant Professor in the Department of Information Technology, Bharathiar University, Tamil Nadu, India. He received his Ph.D. degree in Computer Science from Monomaniam Sundaranar University in the year 2013. M.E., Degree in Computer Science and Engineering from Annamalai University in the year 2007. B.E., Degree in Computer Science and Engineering from Periyar University in the year 2002. He obtained his Diploma in Electronics and Communication Engineering from State Board of Technical Education in the year 1999. He had published over 88 journals papers and over 45 International Journal of Computer Sciences and Engineering Vol.10(5), May 2022, E-ISSN: 2347-2693 © 2022, IJCSE All Rights Reserved 7 conferences papers both at National and International level. His areas of interest include Computer Networks, Network Security, Information Security, etc.

