International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Machine Learning Models for Diabetes Risk Assessment

Pavitha N.[1], Amruta Mankawade[2], Savithramma R. M.[3], Ashwini B. P.[4], Shwetha A. N.[5], Shweta Kambare[6]

Submitted:10/03/2024 **Revised**: 25/04/2024 **Accepted**: 02/05/2024

Abstract: In today's global landscape, diabetes has emerged as a significant and widespread health concern, not limited to India but affecting populations worldwide. Recent years have witnessed the onset of diabetes across all age groups, attributed to various factors such as lifestyle choices, genetic predisposition, stress, and the natural aging process. It is imperative to recognize that any trigger for diabetes can have profound implications if left undetected. In response to this growing health challenge, diverse methodologies are being deployed to predict diabetes and its associated complications. Machine learning algorithms, well-established for predictive analytics across various domains, are gaining prominence in healthcare. Although applying predictive analytics to healthcare is a complex endeavor, it holds the potential to empower healthcare professionals to make informed and timely decisions regarding patient health and treatment options. This study undertakes an investigation into predictive analytics within the healthcare domain, employing six distinct machine learning (ML) algorithms. A comprehensive dataset containing patients' clinical records is employed for testing purposes, and these six diverse ML algorithms are rigorously applied to the dataset.

Keywords: *Diabetes Prediction, Predictive Analytics, Healthcare, Machine Learning Algorithms, Clinical Data, Health Decision Making*

1. Introduction

Diabetes, characterized by insufficient insulin levels in the bloodstream, manifests through symptoms such as irregular urination, excessive thirst, and heightened hunger, all indicative of elevated blood sugar levels [2]. Left untreated, diabetes can precipitate a cascade of health complications, potentially resulting in fatality and multi-organ dysfunction. Two predominant types of diabetes exist: Type 1 and Type 2. The fundamental distinction between them lies in the fact that Type 1 diabetics do not produce insulin, while Type 2 diabetics exhibit reduced responsiveness to insulin. Furthermore, individuals with Type 2 diabetes may progressively produce inadequate insulin levels as the condition advances [4].

In the realm of healthcare, various data mining algorithms have emerged as potent tools for constructing decision support systems. The remarkable precision of these decision-support frameworks underscores their efficacy in aiding healthcare professionals. The overarching objective is the creation of a decision support system capable of reliably predicting and evaluating specific diseases [15]. Artificial Intelligence (AI) encompasses machine learning (ML), a subset of AI that empowers computers to learn autonomously, eliminating the necessity for explicit programming.

Numerous ML algorithms have been introduced to forecast the onset of diabetes. Among these, decision trees stand out as tree-like structures wherein each node represents a feature test, and each leaf node corresponds to a class name. The branches delineate feature combinations leading to class assignments, accommodating both numeric and categorical data [2]. Another ML algorithm, Naive Bayes (NB), leverages the probability of specific outcomes based on independence assumptions among features. This independence assumption underpins precise disease predictions [3]. In contrast, K-Nearest Neighbors (KNN) is characterized as a 'lazy learning' algorithm that employs distinct parameters for classification. Various distance metrics such as Manhattan and Euclidean distance contribute to its independence [4].

This study analyses six distinct classifiers within our proposed framework. The classification model is implemented and meticulously evaluated. Additionally, we assess the outcomes of feature selection and dimensionality reduction experiments, thereby offering effective insights for interventions and enhancements.

2. Literature Review

The research by Veena Vijayan and Anjali C [5] delved into diabetic infections stemming from elevated blood sugar levels. They devised computerized data systems incorporating classifiers such as decision trees, Support Vector Machines (SVM), Naive Bayes, and Artificial

[1] *Department of Artificial Intelligence, Faculty of Science and Technology, Vishwakarma University, Pune, India*
ORCID ID: 0000-0002-0577-8722
** Corresponding Author Email: pavitha.nooji@vupune.ac.in*
[2,6] *Vishwakarma Institute of Technology, Pune, India*
ORCID ID: 0009-0001-8557-514X
ORCID ID: 0009-0007-8957-122X
[3]*Siddaganga Institute of Technology,Tumakuru, Karnataka, India-572103*
ORCID ID: 0000-0002-3176-2329
[4]*Siddaganga Institute of Technology, Tumakuru, Karnataka, India-572103*
ORCID ID: 0000-0001-9511-7292
[5]*Siddaganga Institute of Technology,Tumakuru, Karnataka, India-572103*
ORCID ID: 0000-0003-4981-1464

Neural Networks (ANN) to predict and diagnose diabetes. In [6], the authors undertook predictions concerning diabetes types, complications, and treatment recommendations. Employing predictive analytics and Hadoop framework, they harnessed vast datasets from laboratories, clinics, Electronic Health Records (EHR), and Personal Health Records (PHR) processed via Hadoop, and then distributed the results across diverse servers based on geographical locations.

A comprehensive analysis by Aldo Dagnino and Jiang Zheng [7] explored the application of various machine learning algorithms in the context of power systems. They discussed the implementation of these algorithms in forecasting power system issues, including power grid faults. In [8], a healthcare prediction system centered on the Naive Bayes (NB) algorithm was presented. Utilizing a database of medical measurements, the system extracted hidden information related to various diseases. This framework allowed users to share health-related concerns, subsequently employing NB to predict the most probable ailment.

The optimization of machine learning algorithms for more accurate prediction and analysis of heart disease in continuous disease monitoring was addressed in [3]. The study introduced an advanced convolutional neural network (CNN) emphasizing multimodal approaches for disease detection and risk assessment, leveraging data from a Chinese life healing center between 2013 and 2015. In another experiment focused on chronic cerebral infarction, results demonstrated that for organized data, the Naive Bayes algorithm exhibited superior performance accuracy. Moreover, when combining structured and textual data, the proposed algorithm performed exceptionally well, as evidenced in [10].

The study by Sadegh et al. [11] introduced a framework in the field of data mining for predicting economic events, specifically assessing insolvency and financial hardship using the K-Nearest Neighbors (KNN) algorithm. Marius et al. [12] proposed a framework for selecting appropriate algorithms based on the acquired data, enhancing the speed and accuracy of Nearest Neighbor calculations, particularly in high-dimensional spaces. This approach contributed to precise calculations in computer vision applications.

The framework by Kevin Beyer et al. [13] explored the effects of increasing dimensionality on k-Nearest Neighbors (k-NN) algorithms, emphasizing the diminishing relevance of the difference between the nearest and farthest data points as complexity grows, impacting prediction accuracy. Additionally,Mohamed EL Kourdi et al. [14] introduced a machine learningalgorithm called Credulous Bayes (NB) for organizing Arabic Web archives. They utilized K-Nearest Neighbor (KNN) categorization for predicting financial difficulties and insolvency, a critical concern given the rise in bankrupt companies due to global financial crises. Lloyd's H algorithm and generational algorithms were proposed in Expectation-Maximization (EM) [16]. Decision trees exhibited promise in predicting hyperglycemia with an accuracy of 78.17 percent [17]. Artificial Neural Networks (ANN) and backpropagation algorithms were utilized for pattern recognition and simultaneous classification [17].

3. Materials And Methods

The proposed framework centers on utilizing calculations combinations appeared over within the piece graph. The base classification calculations are Logistic Regression, KNN, Decision Tree, Naive Bayes, and Support Vector Machine for accuracy authentication.

3.1. Dataset Attributes

Pima Indians Diabetes Database is used for the study. There are 768 cases and 9 characteristics in the collection of information data. The primary features of dataset:

- Pregnancy count
- Level of insulin
- Blood pressure in the diastole
- Skinfold thickness in millimeter (mm)
- Body Mass Index (BMI)
- Insulin level 2 hours before and after meal
- Age of patient in years
- Hereditary factor- Pedigree function

For training and validation, the percentage division option exists. Seventy five percent of the seven sixty-eight instances are utilized for teaching, while twenty five percent are used for evaluation [1].

3.2. Methodology

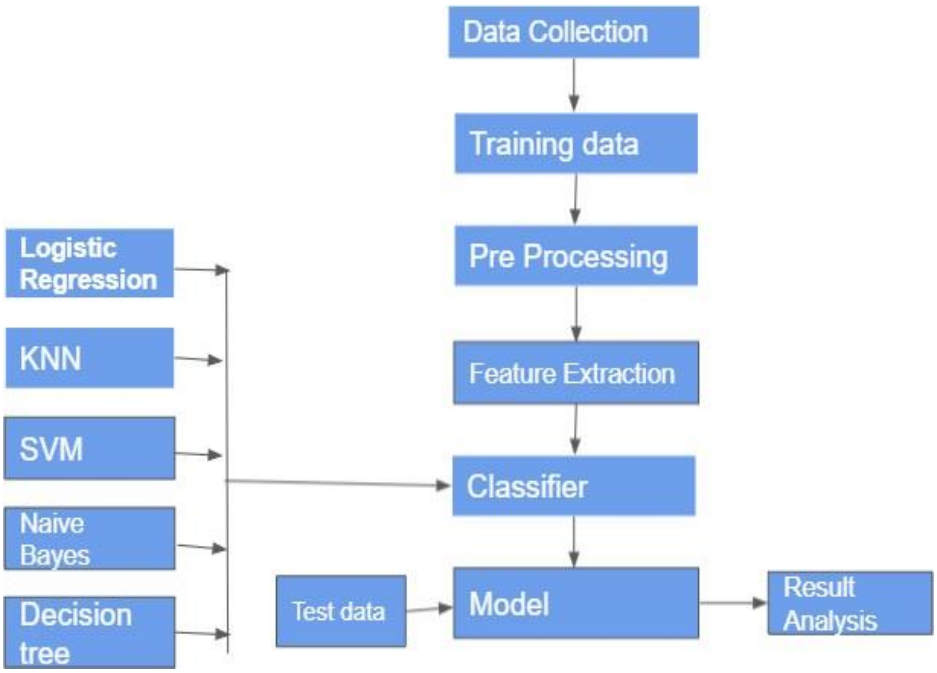

Fig 1.1. Methodology Flow Chart

3.3. Training Data and Testing Data

The machine learning training dataset is used to train the model to perform a variety of actions. For training the model, detailed properties of the training set are gathered. As a result, the design combines these components. Words or sets of successive words are retrieved from tweets using sentiment analysis. [5] As a result, assuming the training set is properly labelled, this model will be able to capture some of the characteristics. This data structure is used to evaluate the model and see if it is performing correctly. [6]

3.4. Pre-processing

Before making the data available to the algorithm some transformations that must be applied on our data. This process is called preprocessing. Most of the time information is collected from distinct sources and gathered in raw format. This Raw data is often not suitable for analyzing directly. The technique of data preprocessing is essential for converting this raw data into understandable data. [3][14]

Data preprocessing is shown below:

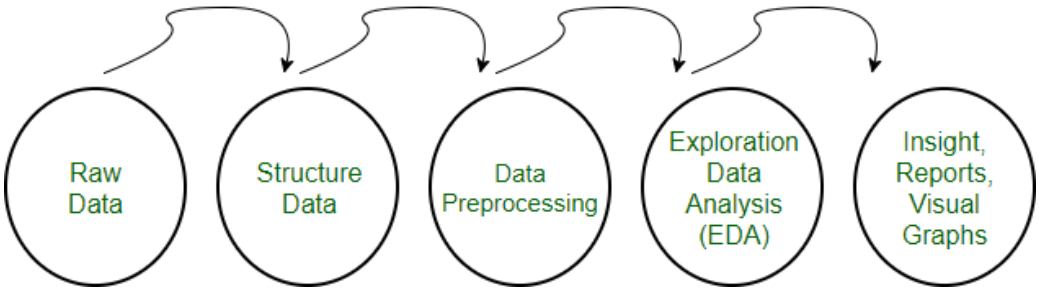

Fig 1.2 Data Preprocessing

3.5. Feature Extraction

Feature Extraction is a technique which is used for converting the input data as the final results of features. Attribute rectangular measures are functions of input designs that facilitate distinguishingamong the instructions of input designs. In case of this algorithm, if the entered statistics is just too large for processing it'll be imagined to be redundant because of the repeat incidence of images which might be represented as pixels, which might be modified right into a condensed set of attributes. Using the extracted characteristic rather than the whole preliminary statistics the selected project may be achieved.[11]

3.6. Machine Learning Algorithms Used

3.6.1. Logistic Regression

Logistic regression (LR) [16] is a widely employed technique for addressing binary classification problems. This method leverages the concept of probabilities to make predictions. While often confused with linear regression, logistic regression employs a more intricate approach involving the sigmoid function, also known as the logistic function. The cost function in logistic regression is composed of two values, 0 and 1, making it apt for binary classification tasks.

3.6.2. K-Nearest Neighbor (KNN) Classification

K-Nearest Neighbor (KNN) [17] is a straightforward yet effective strategy known for its versatility. It operates as an instance-based learning method, combining elements of regression and clustering. KNN is employed to determine the proximity of unlabeled data points to known categories, resulting in highly accurate predictions. It is particularly suitable for initial stages of analysis.

3.6.3. Support Vector Classifier

In comparison to other classifier techniques such as decision trees and logistic regression, Support Vector Machine (SVM) boasts the potential for exceptionally high accuracy. Its ability to handle nonlinear input spaces is a key attribute, making it suitable for applications like image classification, handwriting detection, face detection, and email classification.

3.6.4. Naive Bayes

Naive Bayes (NB) classifiers are a family of simple probabilistic classifiers rooted in Bayes' theorem. They assume strong independence among features and focus on calculating the likelihood of an event occurring based on previous correlated data. NB [19] is renowned for its speed and is particularly well-suited for handling extensive datasets. It excels even when data is limited, owing to its straightforward implementation and direct computation.

3.6.5. Decision Tree

Decision trees [20] are robust modeling tools widely applicable across diverse domains. They are algorithmic constructs that delineate various paths for splitting a dataset based on various conditions. Decision trees are among the most commonly used techniques for supervised learning, aiming to create models that predict the value of a target variable akin to the dependent variable in logistic regression. Decision trees operate on conditional if-then-else statements, making them adept at classification without the need for additional algorithms. They accommodate both continuous and categorical variables.

3.6.6. Random Forest

Random Forest [20][21] is a supervised learning algorithm comprising multiple random decision trees. It represents an evolution of the traditional decision tree method. Random Forest constructs numerous decision trees and combines their outputs to deliver highly accurate predictions with robust forecasts. This versatile and efficient tool is employed to tackle complex learning tasks. By generating an assortment of trees with different feature selections, it mitigates overfitting issues and is often employed towards the conclusion of forecasting tasks.

4. Results and Discussion

For the execution evaluation within the test. To begin, we define TP, FP, TN, and FN as true positive (the probability accurately predicted as required), false positive (the probability inaccurately predicted as required), true negative (the probability accurately predicted as not required), and false negative (the probability inaccurately predicted as not required).Then, we can obtain four measurements: accuracy, precision, recall and F1-measure as follows:

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+TN+FN)}$$

$$\text{Recall} = \frac{TP}{(TP+FP)}$$

$$\text{F1- Measures} = \frac{(2X\ Precision X Recall)}{(Precision+Recall)}$$

TABLE 1.1 Algorithms compared and their performance metrics

Algorithm Implemented	Accuracy	Recall	Precision	F1-Score
Logistic Regression	72.73	76.36	84	80
K Nearest neighbors	77.27	80.37	86	83.10
Support Vector Classifier	72.73	76.36	84	80
Naive Bayes	70.78	75.22	82	78.46
Decision tree	70.78	75.70	81	78.26
Random Forest	75.97	80	84	81.95

In the discussion of the algorithm implementations and their corresponding performance metrics, several

noteworthy observations can be made. Firstly, the K Nearest Neighbors (KNN) algorithm exhibits the highest accuracy at 77.27%, indicating its ability to make correct predictions effectively. It also demonstrates commendable recall, precision, and F1-score values, at 80.37%, 86%, and 83.10%, respectively. This suggests that KNN excels in identifying true positives and minimizing false negatives, making it a strong candidate for this classification task.

Logistic Regression and Support Vector Classifier (SVC) both achieve an accuracy of 72.73%, accompanied by respectable recall, precision, and F1-scores at 76.36%, 84%, and 80%, respectively. These algorithms provide a balanced trade-off between sensitivity and precision, demonstrating their utility in predictive modeling for this context. On the other hand, Naive Bayes, Decision Tree, and Random Forest algorithms all yield relatively similar accuracy scores in the range of 70.78% to 75.97%. These models exhibit reasonable recall values but slightly lower precision, indicating a tendency to generate some false positives. However, the F1-scores for these algorithms are reasonably high, suggesting a balance between recall and precision.

In conclusion, while K Nearest Neighbors emerges as the top-performing algorithm in terms of accuracy and overall predictive capability, other algorithms like Logistic Regression, Support Vector Classifier, Naive Bayes, Decision Tree, and Random Forest also offer viable options depending on the specific requirements of the application. The choice of algorithm should be driven by the desired balance between sensitivity and precision, taking into account the context and objectives of the predictive modeling task. Further optimization and fine-tuning of these algorithms may enhance their performance and applicability in real-world scenarios.

Conclusion

Upon conducting a meticulous review of the introductory overview, it becomes evident that the earlier expectations were founded on a relatively limited dataset. It is well-established that the utilization of a more extensive dataset significantly enhances the predictive capabilities of any system. The architecture we are developing holds immense potential for application in the medical sector, particularly for the management and monitoring of treatment checkups for individuals with diabetes mellitus. Our research endeavors aim to harness the power of comprehensive data and advanced algorithms to create a framework that not only enhances prediction accuracy but also promises to be a valuable asset in the realm of healthcare, contributing to improved patient care and well-being.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Shetty, D., Rit, K., Shaikh, S., & Patil, N. (2017). Diabetes disease prediction using data mining. 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS). doi:10.1109/iciiecs.2017.8276012

[2] K.Suresh, O.Obulesu, & B. Venkata Ramudu.(2020). Diabetes Prediction using Machine Learning Techniques. Helix, 10(02), 136-142.

[3] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," IEEE Access, vol. 5, pp. 8869-8879, 2017.

[4] Jakka, Aishwarya & Jakka, Vakula. (2019). Performance Evaluation of Machine Learning Models for Diabetes Prediction. 10.35940/ijitee.K2155.0981119.

[5] Veena Vijayan V. And Anjali C, Prediction and Diagnosis of Diabetes Mellitus, "A Machine Learning Approach" ,2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) | 10- 12 December 2015 | Trivandrum.

[6] N. M. S. Kumar, T. Eswari, P. Sampath, and S. Lavanya, "Predictive Methodology for Diabetic Data Analysis in Big Data," Procedia Comput. Sci., vol. 50, pp. 203–208, Jan. 2015.

[7] J. Zheng and A. Dagnino, "An initial study of predictive machine learning analytics on large volumes of historical data for power system applications," in 2014 IEEE International Conference on Big Data (Big Data), 2014, pp. 952–959.

[8] International Journal of Advanced Computer and Mathematical Sciences. Bi Publication- BioIT Journals, 2010.

[9] R. A. Taylor et al., "Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big DataDriven, Machine Learning Approach," Acad. Emerg. Med., vol. 23, no. 3, pp. 269–278, Mar. 2016

[10] Sadegh B.Imandoust, M.Bolandraftar, "Application of KNearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background",International Journal of Engineering Research and Applications, Vol. 3, 2013.

[11] M.Muja, David G.Lowe, "Fast Approximate Nearest Neighbors With Automatic Algorithm Configuration", University of British Columbia

[12] K Beyer, J Goldstein,R Ramakrishnan and U Shaft, "When is 'Nearest neighbor' Meaningful?" 2014

[13] M.Elkourdi, A.Bensaid, T.Rachidi, “Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm”, Alakhawayn University, 200

[14] Velu C.M, K. R. Kashwan, "Visual Data Mining Techniques for Classification of Diabetic Patients”, IEEE International Advance Computing Conference (IACC), pp-1070-1075,2013

[15] Asma A. AlJarullah, "Decision discovery for the diagnosis of Type II Diabetes", IEEE conference technology.pp-303-307,2011. on innovations in information

[16] Peng, Joanne & Lee, Kuk & Ingersoll, Gary. (2002). An Introduction to Logistic Regression Analysis and Reporting. Journal of Educational Research - J EDUC RES. 96. 3-14. 10.1080/00220670209598786.

[17] Guo, Gongde& Wang, Hui & Bell, David & Bi, Yaxin. (2004). KNN Model-Based Approach in Classification.

[18] Evgeniou, Theodoros & Pontil, Massimiliano. (2001). Support Vector Machines: Theory and Applications. 2049. 249-257. 10.1007/3-540-44673-7_12.

[19] Rish, Irina. (2001). An Empirical Study of the Naïve Bayes Classifier. IJCAI 2001 Work Empir Methods ArtifIntell. 3.

[20] Patel, Harsh & Prajapati, Purvi. (2018). Study and Analysis of Decision Tree Based Classification Algorithms. International Journal of Computer Sciences and Engineering. 6. 74-78. 10.26438/ijcse/v6i10.7478.

[21] Ali, Jehad & Khan, Rehanullah& Ahmad, Nasir & Maqsood, Imran. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues(IJCSI).