Quest Journals
Journal of Software Engineering and Simulation
Volume 10 ~ Issue 7 (2024) pp: 08-13
ISSN(Online) :2321-3795 ISSN (Print):2321-3809
www.questjournals.org

Research Paper

Diabetes Prediction Using SVM and Logistic Regression Techniques

Deepika Singh[1], Shubham Singh[2]
Scholar M.Tech in Computer Science & Engg. (GNIOT, Gr. Noida) [1]
Engineer in Max Healthcare Institute Ltd.[2]
singhdeepika3469@gmail.com [1], shubhamtitm@gmail.com[2]

Abstract.
The most alarming diseases are affecting individuals worldwide. To help prevent and manage these conditions, a machine learning model requires time to inform people about their health status. Many existing machine learning models focus on specific diseases. This research introduces a robust and efficient machine learning method to detect the three most prevalent health conditions: diabetes, heart disease, and COVID-19. The model was evaluated using KNN, Logistic Regression, and Support Vector Machine techniques. The model's effectiveness is determined by the number of parameters it analyzes. The criteria used to diagnose diabetes include pregnancies & blood pressure, insulin levels, skin breadth, BMI, diabetes full-blooded function, age, and conclusion. Additional metrics are considered for analyzing heart conditions and Covid-19. The experimental results show accuracy rates of 77.64% and 77.34% for diabetes diagnosis, 87.00% for heart disease detection, and 90.00% for identifying Covid-19 disorders.

Keywords: *Diabetes, Heart, SVM, Logistic Regression, Accuracy*

I. Introduction

In recent times, chronic illnesses have been increasingly affecting populations worldwide, in both wealthy and impoverished countries. Among these chronic conditions, diabetes mellitus stands out for its significant impact on individuals' well-being from an early age. The surge in diabetes cases is primarily driven by rising obesity rates and an aging population. To address this, it is crucial to assess the high-risk groups for diabetes mellitus (DM) using modern information technology tools

Data mining, or Knowledge Discovery in Databases (KDD), involves using computational methods to uncover patterns within extensive datasets, drawing on techniques from Artificial Intelligence, Machine Learning, Statistics, and Database Systems. The prime goals of these techniques contain identifying patterns, making predictions, establishing associations, and performing clustering tasks. Essentially, data mining uses various automated or semi-automated approaches to extract and identify valuable, previously unknown information from large amounts of data.

The principles of statistics mining emphasize importance of high-quality data & appropriate technology [3]. Health care organizations can reduce costs and improve service quality through the use of data mining. Predictive analysis has been utilized to determine whether an article has diabetes. A multiplicity of computational methods have been employed to classify diabetes in patients. The integration of machine learning into medical information systems has proven beneficial for diabetes treatment by enhancing the number of available treatments, reducing costs, and improving diagnostic accuracy. Diabetes is a type of metabolic disorder that is characterized Diabetes is characterized by increased blood glucose levels due to issues with insulin activity, emission, or mutually. Persistent hyperglycemia in Diabetes can lead to chronic damage, irregularities, and organ malfunction over time, such as the kidneys, eyes, nerves, heart, and blood vessels. The pathogenic mechanisms for diabetes to spread are numerous.

The heart is a crucial organ in the human body, functioning as a muscle divided into two sections by valves, creating four chambers. Each section consists of an atrium and a ventricle. Blood first enters the atria (the plural form of atrium), and then the ventricles contract to pump it out of the heart. The right side of the heart pumps oxygen-poor blood to the lungs, where it receives oxygen. This newly oxygenated blood is then sent to the left atrium and left ventricle. Finally, the left ventricle pumps oxygen-rich blood. throughout body's tissues and organs.

Oxygenated blood plays a crucial role in enhancing energy levels and overall health. According to the National Heart, Lung, and Blood Institute (2008), the term "heart disease" encompasses various conditions affecting the heart and blood vessels. Heart disease causes millions of myocardial infarctions in both men and women across the United States. Symptoms of heart disease can differ based on the specific type of condition, with chest pain being a common indicator. Atherosclerosis is a particularly insidious form of heart disease, as some individuals might not exhibit any symptoms until the condition becomes life-threatening. Regardless of the various forms heart disease can take, everyone at risk shares common risk factors. Key factors include blood pressure (BP), age, and gender.

COVID-19 is a significant global pandemic with approximately 550,000 reported cases, resulting in nearly 25,000 deaths. This virus is part of the coronavirus family, which can infect both humans and animals. The outbreak began in late 2019 in Wuhan, a city of 11 million people in China. Over the last two decades, two other noteworthy coronavirus outbreaks have been the Severe Acute Respiratory Syndrome (SARS) in 2003 and the Middle East Respiratory Syndrome (MERS) in 2012. Common symptoms of COVID-19 include fatigue, body aches, nasal congestion, fever, and a dry cough.

The structure of the document is prepared as follows: Section 2 provides an analysis of the text on the disease. prediction system, Section 3 introduces the machine learning techniques employed in the study, Section 4 presents the conclusion and examination, and Section 5 discuss the conclusions drawn and outlines future directions for research.

II. Literature Survey

A lot of work has been done for the forecast of diseases via machine learning technique in the past.Some of the best literature work is presented here. Asgarnezhad et al., [7] included the combination of attribute subset selection techniques and mislaid value substitution for presenting an effective pre-processing technique. The most popular diabetes mellitus data set is used for experimentation. The applied classifier performance can be enhanced using this projected technique as shown in investigational results and diabetes mellitus prediction, with respect to precision and accuracy, traditional techniques are outperformed using this technique.

Maulana and Endah [8] compared Stepwise Regression (SR), Forward Selection (FS), and Backward Elimination (BE) algorithms with respect to attribute selection for implementing an effective data preprocessing method and K-Nearest Neighbor Algorithm are used for diabetic data classification. Medical record data has been collected from diabetes disease information management of Pusat Pertamina Hospital and the tests were conducted using the K-Fold Cross Validation strategy (k=10). There are 126029 data. In diagnosing diabetes, accuracy can be enhanced using data preprocessing stages in aspect selection, as illustrated in the outcome. On pre-processing data, Stepwise Regression Algorithm produces better accuracy.

Wosiak and Karbowiak [9] classified imbalanced medical datasets in an effective way using a developed pre-processing compensation method. For some datasets, there is a combination of classification and certain preprocessing technique, which outperforms other techniques. Irrespective of the applied classification method and re-sampling technique, there will be a reduction of correctly predicted labels in datasets having large features or complex distribution.

Bai et al., [10] presented a pre-processing method in the mining process while working on medical datasets. Missing values in the medical dataset are handled using data mining, which is included in the pre-processing stage. In a medical dataset with categorical attribute values, issues in missing values handling are addressed in this work. Missing values are estimated and fixed using a proposed imputation measure in this research work.

Walczak et al., [11] obtained enhanced computer diagnosis accuracy using an implemented framework in medical patient checkups. According to medical data pre-processing, constructed a new proposition for analysis of medical data. Medical data are transformed to parameterized mathematical formulations from a descriptive, semantic form using this pre-processing. Moreover, it provides an effective framework for revealing concealed features within medical data. Pre-processing enables new possibilities for interpreting medical data by examining these hidden characteristics. Medical databases utilize parameterized illness patterns to enhance diagnostic accuracy.

Wu et al. [12] developed innovative algorithms based on data mining to predict Type 2 Diabetes Mellitus (T2DM). Their approach involves a series of pre-processing techniques that split the model into two parts: enhanced K-Means and logistic deterioration algorithms. They used the Waikato Environment for Knowledge Analysis (WEKA) toolset to conduct experiments and compare results with the Pima Indians Diabetes Dataset. Their findings show that the proposed model improves accuracy by approximately 3.04% compared to previous methods, demonstrating that the dataset quality is well-maintained. These results highlight the effectiveness of the proposed methodology for practical diabetes health management.

Pradhan and Bamnote [13] have developed an effective binary classifier for diabetes detection that employs Support Vector Machine (SVM) techniques and data preprocessing. This research focuses on reducing the number of features by applying an attribute evaluator combined with a best-first search algorithm, resulting in the use of three features instead of the original eight. The experiments are conducted using the Pima Indian Diabetes dataset from the UCI repository. Data preprocessing plays a crucial role in significantly enhancing the model's accuracy.

Raihan et al. (2016) enhanced a simple concept to predict the rising risk of Ischemic Heart Disease (IHD) using smartphones. They integrated clinical data from IHD-diagnosed patients into an Android application to improve its accuracy. This research aims to build on this idea, allowing individuals to assess their IHD risk and encouraging them to consult a cardiologist proactively to prevent sudden cardiac death. The availability of devices was restricted to prevent misuse. The proposed system could aid in timely risk assessment and help mitigate limitations.

Alqaraawi et al. [15] Proposed an adaptive instantaneous concept to employ Linear Predictive Coding (LPC) and Wavelet transformation methods to estimate Heart rate variability (HRV) from Photo PlethysmoGraphy (PPG) signal documented by wearable devices. The projected algorithm had performed well on two associated algorithms, particularly for low PPG signal to noise ratio. By evaluating this projected algorithm to the ground certainty accounted concurrently from ECG, an average temporal resolution of 8.7 ms had attained with a 82.9% sensitivity and a 82.7% positive predictive value.

Mufudza & Erol [16] discussed the pattern based clustering methods for prediction and to make a diagnosis of heart disease through poison combination decay. A Zero overstated poison combination decay model had been ejected to be the optimal pattern to predict heart disease. It was presumed that heart disease prediction had been efficiently carried out by detecting the significant risks component using Poisson Mixture Regression model.

B. Jena [17] introduced the diagnosis system for disease prediction in healthcare. The ANN machine learning approach is implemented on Cleveland dataset. The highest value for correctly classifying the HD is 83% for 5-fold cross validation. The merits of ANN approach is utilized in predicting the disease in starting stage which helps physicians to take care of the patient’s life.

Zeinab Arabasadi et al [18] examined that cardiovascular infection is a standout amongst the most uncontrolled reasons for death around the globe and was considered as a noteworthy sickness in middle and old ages. Coronary heart sickness, specifically, is an across the board cardiovascular disease involving high death rates. In this way, much research has been led utilizing machine learning and data mining algorithms in order to look for elective modalities.

III. Machine Learning Techniques

3.1 **Support Vector Machine (SVM)**

Support Vector Machine is a binary classifier, which is based on maximum margin separation principle. The first term represents the regularization, which is given by the normal of separating plane w. Subsequently in the second term, the empirical risk is evaluated for training the datasets and is weighted by the parameter x. The Hinge loss function (L) is defined for the soft-margin classifier as,

$$\mathrm{L} = \max(0, 1 - \mathrm{yiwTxi}) \quad (1)$$

3.2 Logistic Regression

Logistic regression uses logit function in its basic form to model binary response variable. Logistic regression is the appropriate model when there are two possible outcomes, often described as a binary response. In some contexts, one outcome is designated as a success, while the other is considered a failure. The logistic regression model estimates the probability of success rather than simply classifying an outcome as success or failure. To handle this, the response variable is transformed into an odds ratio, p / (1-p), which spans the positive real numbers, since the probability of success is confined to values between 0 and 1.To guarantee that the response variable is a real number, the odds ratio is adjusted using the natural logarithm function ln(p/(1-p)), which ensures that the predicted values fall within the acceptable range. After this transformation of the response variable, the subsequent steps mirror those used in linear regression. The logistic regression equation is given by:

f(w) = linear regression function where,

$$\mathrm{f(w)} = \gamma 0 + \gamma 1\, \mathrm{w1} + \gamma 2\, \mathrm{w2} + \cdots .. + \gamma \mathrm{p}\, \mathrm{wp} + \varepsilon. \quad (2)$$

3.3 K-Nearest Neighbor

The k-Nearest Neighbors (kNN) algorithm is a straightforward and often effective non-parametric classification technique. To classify a data record \(t \), a neighborhood is established by identifying its \(k \) closest neighbors. Typically, the classification of \(t \) is determined by a majority vote among these neighboring data records, which can be weighted by distance or not. However, selecting an appropriate value for \(k \) is crucial, as the classification outcome is significantly influenced by this choice. The kNN method can be affected by the value of \(k \). One common approach to determining \(k \) is to iteratively test the algorithm with various \(k \) values and select the one that performs the best.

To understand the detailed functioning of the algorithm, the following steps should be undertaken:

Step 1: Given the training dataset:

{ (x(1), y(1)) , (x(2), y(2)), …… , (x(m), y(m)) }

Step 2: hoard the trained data

Step 3: For each data point that is unlabeled:

(a) Euclidean distance is computed for every data point.

(b) k- nearest neighbors are tracked

(c) Class assignment for cluster having maximum neighbors.

(d) Normalize all the parameters to make calculation easier.

(e) Select the K-values at different thresholds to find the best accuracy.

IV. Experimental Results and Discussions

Experimental results are carried out using python language with anaconda platform. The hardware requirement used for the result analysis are as follows: Intel Corei3, 6GB RAM with CPU speed of 2.12HZ.

(a) Diabetes Analysis using Support Vector Machine

For this research, the samples of over 2000 people have been collected randomly from different regions. The parameters on which the research is conducted to conclude whether the person is diabetic or not are: No of PregnanciesGlucose level, Blood Pressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age. The input data is subjected to histogram equalization process and then correlation is determined by removing outliers from the database. Then the database is trained by bifurcating into 80:20 ratio. Then, after building the confusion matrix and applying cross validation technique, the system delivers the accuracy rate of 77.64%.

(b) Diabetes Analysis using Logistic Regression

For this research, the samples of over 2000 people have been collected randomly from different regions. The parameters on which the research is conducted to conclude whether the person is diabetic or not are: Sure, here's a rewritten version Factors such as blood pressure, age, skin thickness, insulin levels, BMI (body mass index), diabetes family history, glucose levels, and the number of pregnancies are all important considerations. The input data is subjected to histogram equalization process and then correlation is determined by removing outliers from the database. Then the database is trained by bifurcating into 80:20 ratio. Then, after building the confusion matrix and applying cross validation technique, the system delivers the accuracy rate of 77.34%.

Figure 2 compares the accuracy of six machine learning algorithms. Adaboost achieves the highest accuracy. Random forest and logistic regression follow, both displaying similar accuracy but greater precision. SVM also performs well with an accuracy of 82.46%. The accuracies of decision trees and xgboost are both below 80%, with xgboost being the least accurate among the six.

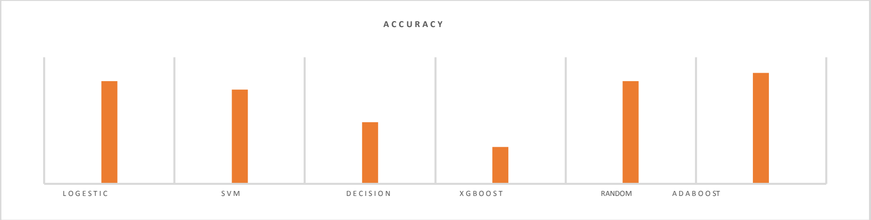

Fig. 2. Comparison of the Accuracy-Based Machine Learning Algorithms' Performance

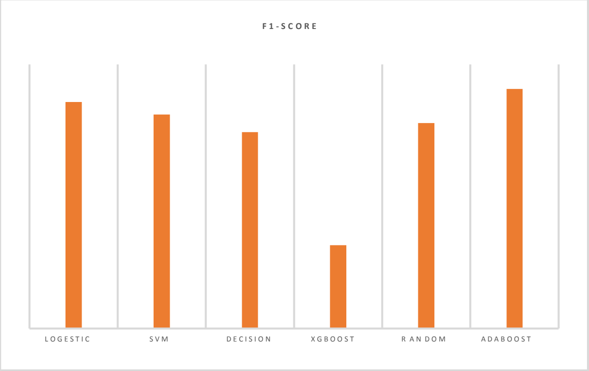

Fig. 3. Machine Learning Algorithm Performance Comparison Based on F1-score

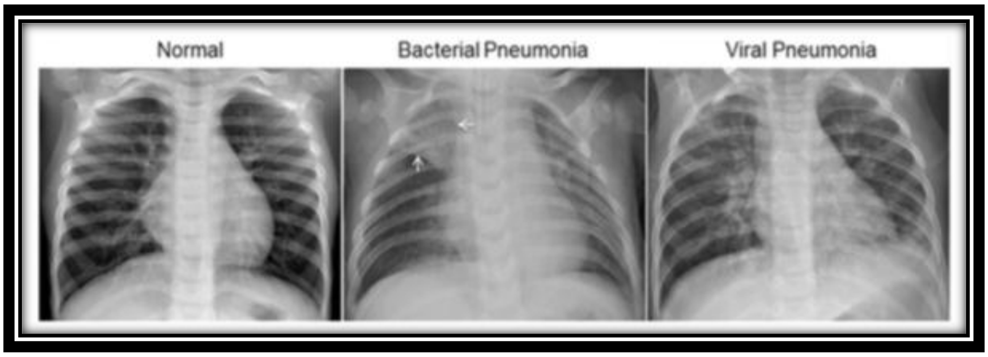

Fig. 3 Chest X-Ray dataset images from Kaggle

Table 1. Accuracy Analysis of diseases using machine learning techniques

Name of Diseases	Algorithm used for prediction	Accuracy Rate
Diabetes	Support Vector Machine	77.4%
	Logistic Regression	77.34%

V. Conclusion and Future Scope

Machine learning algorithms have been worn extensively for prediction of diseases. In this investigate; several machine learning algorithms are practical to predict diabetes percentage. The advantage of using this technique lies on the less computational value and availability of data. The accuracy rate for diabetes is 77.40% using SVM and 77.34% using Logistic Regression. The future scope in this area can be to apply machine learning techniques on big data for further research.

References

[1]. Alehegn, M., and Joshi, R., Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. International Research Journal of Engineering and Technology, Vol. 4(10), pp. 426-435, 2017.

[2]. Kaul, K., Tarr, J. M., Ahmad, S. I., Kohner, E.M. and Chibber, R., Introduction to diabetes mellitus. Adv. Exp. Med Biol. 2012;771:1- 11. doi:10.1007/978-1-4614-5441-0_1.

[3]. Lingaraj, H., Devadass, R., Gopi, V. and Palanisamy, K., Prediction of diabetes mellitus using data mining techniques: a review. Journal of Bioinformatics & Cheminformatics, Vol. 1(1), pp. 1-3, 2015.

[4]. Rajesh, K. and Sangeetha, V., Application of data mining methods and techniques for diabetes diagnosis. International Journal of Engineering and Innovative Technology, Vol. 2(3), pp. 224-229, 2012.

[5]. Perveen, S., Shahbaz, M., Guergachi, A. and Keshavjee, K., Performance analysis of data mining classification techniques to predict diabetes. Procedia Computer Science, Vol. 82, pp. 115-121, 2016.

[6]. Han, J., Rodriguez, J. and Beheshti, M., Diabetes Data Analysis and Prediction Model Discovery Using Rapid Miner. Second International Conference on Future Generation Communication and Networking, Vol. 3, 96-99 (2008).

[7]. Asgarnezhads, R., Shekofteh, M. and Boroujeni, F.Z., Improving diagnosis of diabetes mellitus using combination of preprocessing techniques. Journal of Theoretical & Applied Information Technology, Vol. 95(13), pp.2889-2895, 2017.

[8]. Maulana, F. and Endah, S.N., Comparison selection of attributes in preprocessing data for diagnosis of diabetes. IEEE International Conference on Informatics and Computational Sciences, pp. 141-146, 2017.

[9]. Wosiak, A. and Karbowiak, S., Preprocessing compensation techniques for improved classification of imbalanced medical datasets. IEEE Federated Conference on Computer Science and Information Systems, Vol. 11, pp. 203-211, 2017

[10]. Bai, B. M., Mangathayaru, N. and Rani, B. P., An approach to find missing values in medical datasets. ACM Proceedings of the

International Conference on Engineering & MIS 2015, doi: 10.1145/2832987.2833083.

[11]. Walczak, A. and Paczkowski, M., Medical data preprocessing for increased selectivity of diagnosis. Bio-Algorithms and MedSystems, Vol. 12(1), pp. 39-43, 2016

[12]. Wu, H., Yang, S., Huang, Z., He, J. and Wang, X., Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked, Vol. 10, pp. 100-107, 2018.

[13]. Pradhan, M. and Bamnote, G. R., Efficient binary classifier for prediction of diabetes using data preprocessing and support vector machine. In Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications, pp. 131- 140, 2015.

[14]. Raihan, M, Mondal, S, More, A, Sagor, M. O. F, Sikder, G, Majumder, MA & Ghosh, K 2016,"Smartphone based ischemic heart disease (heart attack) risk prediction using 111 clinical data and data mining approaches, a prototype design", Computer and Information Technology (ICCIT), 19th International Conference on IEEE, pp. 299-303.

[15]. Alqaraawi, A, Alwosheel, A &Alasaad, A 2016, "Towards efficient heart rate variability estimation in artifact-induced Photo plethysmography signals", Electrical and Computer Engineering (CCECE), IEEE Canadian Conference on IEEE, pp. 1-6.

[16]. Mufudza, C & Erol, H 2016, "Poisson Mixture Regression Models for Heart Disease Prediction", Computational and Mathematical Methods in Medicine.

[17]. Tarle, Balasaheb, and Sudarson Jena, 2017, "An artificial neural network based pattern classification algorithm for diagnosis of heart disease", International Conference on Computing, Communication, Control and Automation (ICCUBEA), pp. 1-4.IEEE.

[18]. Zeinab Arabasadi&RoohallahAlizadehsani 2017, "Computer aided decision making for heart disease detection using hybrid neural network Genetic algorithm", Computer Methods and Programs in Biomedicine, Vol. 141, No. 1, pp. 19-26.