Materials Today: Proceedings xxx (xxxx) xxx

Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

Multiple disease prediction using Machine learning algorithms

K. Arumugam [a], Mohd Naved [b], Priyanka P. Shinde [c,*], Orlando Leiva-Chauca [d], Antonio Huaman-Osorio [e], Tatiana Gonzales-Yanac [d]

[a] Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, Tamilnadu, India
[b] Department of Business Analytics, Jagannath University, Delhi-NCR, India
[c] Master of Computer Application Department, Government College of Engineering, Karad 415124, Maharashtra, India
[d] Administration and Tourism Faculty, Universidad Nacional Santiago Antúnez de Mayolo, Huaraz, Perú
[e] Economics and Accounting Faculty, Universidad Nacional Santiago Antúnez de Mayolo, Huaraz, Perú

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Data mining
Machine learning
Decision tree
Naïve bayes
Support vector machine
Accuracy
Classification
Prediction

ABSTRACT

Data mining for healthcare is an interdisciplinary field of study that originated in database statistics and is useful in examining the effectiveness of medical therapies. Machine learning and data visualization Diabetes-related heart disease is a kind of heart disease that affects diabetics. Diabetes is a chronic condition that occurs when the pancreas fails to produce enough insulin or when the body fails to properly use the insulin that is produced. Heart disease, often known as cardiovascular disease, refers to a set of conditions that affect the heart or blood vessels. Despite the fact that various data mining classification algorithms exist for predicting heart disease, there is inadequate data for predicting heart disease in a diabetic individual. Because the decision tree model consistently beat the naive Bayes and support vector machine models, we fine-tuned it for best performance in forecasting the likelihood of heart disease in diabetes individuals.

Selection and peer-review under responsibility of the scientific committee of the International Conference on Nanoelectronics, Nanophotonics, Nanomaterials, Nanobioscience & Nanotechnology.

1. Introduction

In terms of data collecting and processing, healthcare is one of the most worrisome industries. With the advent of the digital era and technological advancements, a vast quantity of multidimensional data on patients is created, including clinical factors, hospital resources, illness diagnostic information, patients' records, and medical equipment. The enormous, dense, and complex data must be processed and evaluated in order to extract knowledge for effective decision making. Medical data mining offers a lot of potential for uncovering hidden patterns in medical data sets [1].

By identifying significant patterns and detecting correlations and relationships among many variables in huge databases, the use of various data mining tools and machine learning approaches has changed healthcare organizations [2,3]. It serves as an important instrument in the medical sector, providing and comparing existing data for the future course of action. This technology combines multiple analytic methodologies with modern and complex algorithms, allowing for the exploration of massive amounts of data [4]. It is used in healthcare to gather, organize, and analyze patient data in a systematic manner. It may be used to identify inherent inefficiencies and best practices for providing better services, which may lead to improved diagnosis, better medicine, and more successful treatment, as well as a platform for a deeper knowledge of the mechanisms in practically all elements of the medical domain. Overall, it assists in the early detection and prevention of disease epidemics by searching medical databases for pertinent information.

The process of determining a condition based on a person's symptoms and indicators is known as medical diagnosis. In the diagnostic process, one or more diagnostic procedures, such as diagnostic tests, are performed. Diagnosis of chronic illnesses is a vital issue in the medical industry since it is based on many symptoms. It is a complex procedure that frequently leads to incorrect assumptions. When diagnosing illnesses, the clinical judgment is based mostly on the patient's symptoms as well as the physicians' knowledge and experience [5]. Furthermore, when medical sys-

* Corresponding author.
E-mail addresses: priyanka.shinde@gcekarad.ac.in (P.P. Shinde), oleivac@unasam.edu.pe (O. Leiva-Chauca), ahuamano@unasam.edu.pe (A. Huaman-Osorio), dgonzalesy@unasam.edu.pe (T. Gonzales-Yanac).
URL: https://orcid.org/0000-0001-8183-5402 (O. Leiva-Chauca).

https://doi.org/10.1016/j.matpr.2021.07.361
2214-7853/
Selection and peer-review under responsibility of the scientific committee of the International Conference on Nanoelectronics, Nanophotonics, Nanomaterials, Nanobioscience & Nanotechnology.

tems evolve and new treatments become available, it becomes more difficult for physicians and doctors to stay up with the current innovations in clinical practice [6]. For effective therapy, medical practitioners and doctors must be well-versed in all pertinent diagnostic criteria, patient history, and a mix of medication therapy. However, mistakes are possible since they make judgments instinctively based on information and experience gained from past experience with patients. Because of factors such as multitasking, restricted analysis, and memory capacity, their cognitive capacities are restricted [7]. As a result, it is difficult for a physician to make the right judgment on a consistent basis if he is not supported by clinical tests and patient history information. Even experienced physicians can benefit from a computer-aided diagnostic system in making sound medical judgments [8]. Thus, medical professionals are very interested in automating the diagnosis process by integrating machine learning techniques with physician expertise [9]. Data mining and machine learning approaches are making significant efforts to intelligently translate accessible data into valuable information in order to improve the diagnostic process's efficiency. Several studies have been conducted to explore the use of machine learning in terms of diagnostic abilities. It was discovered that, when compared to the most experienced physician, who can diagnose with 79.97% accuracy, machine learning algorithms could identify with 91.1% correctness [10]. Machine learning techniques are explicitly used to illness datasets to extract features for optimal illness diagnosis, prediction, prevention, and therapy.

2. Related work

A structural model and a collection of conditional probabilities are used by Bayesian classifiers. They make the assumption that the contributions of all factors are independent. It first calculates the prior probability for each class, and then applies the occurrence of each variable value to an unknown scenario. A Bayes network classifier is built on a Bayesian network, which reflects a joint probability distribution over a set of category characteristics.

The SVM method and the Nave Bayes technique were used to predict kidney disease [11]. The authors attempted to categorize various stages of kidney disease using the suggested ANFIS algorithm. The study's purpose was to design an effective categorization algorithm using several assessment metrics such as accuracy and execution time. While the SVM Algorithm provided higher classification accuracy, the Nave Bayes fared better since it produced results in less time. The results show that SVM outperforms the Nave Bayes Approach in predicting renal illness.

The fuzzy technique with a membership function was used to forecast cardiac disease [12]. Using the Fuzzy KNN Classifier, the authors attempted to eliminate ambiguity and uncertainty from data. The 550-record dataset was separated into 25 classes, with each class having 22 items. The dataset was separated into two equal parts: training and testing. The fuzzy KNN methodology was implemented after pre-processing techniques were used. This technique was examined using several assessment metrics such as accuracy, precision, and recall, among others. Based on the data, it was discovered that the fuzzy KNN classifier outperformed the KNN classifier in terms of accuracies.

For the prediction of cardiac disease, a novel technique based on the ANN algorithm was devised [13]. The researchers created an interactive prediction method based on categorization using an artificial neural network algorithm and taking into account the thirteen most important clinical parameters. The suggested method proved effective for predicting heart disease with an accuracy of 80% and can be very useful for healthcare practitioners.

Authors in [14] presented an automated approach for answering difficult inquiries for heart disease prediction. The Naive Bayes methodology was used to create this intelligent system in order to provide quick, better, and more accurate outcomes. It might aid doctors in making clinical judgments about heart attacks. This system may be enhanced by including SMS functionality, building Android and IOS mobile applications, and including a pacemaker in the order.

Diabetes and breast cancer were diagnosed by incorporating the adaptivity characteristic into support vector machines [15]. The goal was to offer a rapid, automated, and adaptable diagnostic method using adaptive SVM. To achieve better results, the bias value in conventional SVM was changed. The suggested classifier produced output in the form of 'if-then' rules. The proposed method was used to diagnose diabetes and breast cancer, and it provided 100% right classification rates for both conditions. Future research should focus on developing more efficient ways for changing the bias value in conventional SVM.

For the prediction of type 2 diabetes, a hybrid model based on clustering followed by classification was proposed [16]. For prediction, the suggested model uses K-means clustering and the C4.5 classification method with k-fold cross-validation. The model generated encouraging results with a classification accuracy of 88.38 percent using the hybrid technique, which might be highly useful for clinicians in making appropriate clinical choices related to diabetes.

3. Framework for multiple disease prediction

In this framework, machine learning algorithms- support vector machine, naïve bayes, decision tree are used.

The Naive Slogan The Bayes classification [14] refers to a fundamental probabilistic classification based on strong independent assumptions in the application of the Bayes theorem. The existence or absence of a particular class feature does not depend on the presence or absence of any other feature. It operates on the basis of conditions. It uses Bayes' theorem that determines the probability that an event happens when another event happens. If B represents the dependent event and A represents the last event the theorem Bayes may be phrased as follows: Sample (B supplied in A) = Sample (A and B)/Sample (A and B) (A) . The approach divides the number of events in which A and B occur together by the number of circumstances in which A occurs to get the likelihood of B given A alone. In order to estimate the parameters (variable media and variances), the Naive Bayes Classifier benefits from only a few training data. Due to the assumption of independent variables, all the variances must be computed for each class. It is relevant to binary as well as multi-class problems.

SVM [15] is a method often used for kernel learning to handle issues of large prediction. The SVM classifier has shown greater generalization and a well-scaling of both linear and nonlinear data as compared to other classifiers. In addition, the SVM classificator delivers very strong pattern recognition performance in conjunction with various frequently used approaches in statistical learning and optimisation theory. Identifying an overview that separates positive examples from negative data with the greatest error margin is the main aim of the SVM classification system.

When the data is linearly separable, it is easy to choose the optimum hyper-plane splitting two classes of data. For non-inlinear mapping to large dimension space for non-separable problems, SVM applies 'Kernel Functions' on the other side. There are a number of kernels functions including Linear Kernel Function (LKF), Polynomial Kernel Function (PKF) and Sigmoid Kernel Function (SKF), Exponential Radial Basis Kernel Function (ERBKF) (GRBKF). The Radial Basic Function (RBF) has been identified as the finest kernel function among the several kernel functions.

Decision trees are used [16] extensively for categorizing huge datasets. Decision trees categorize data between the root node

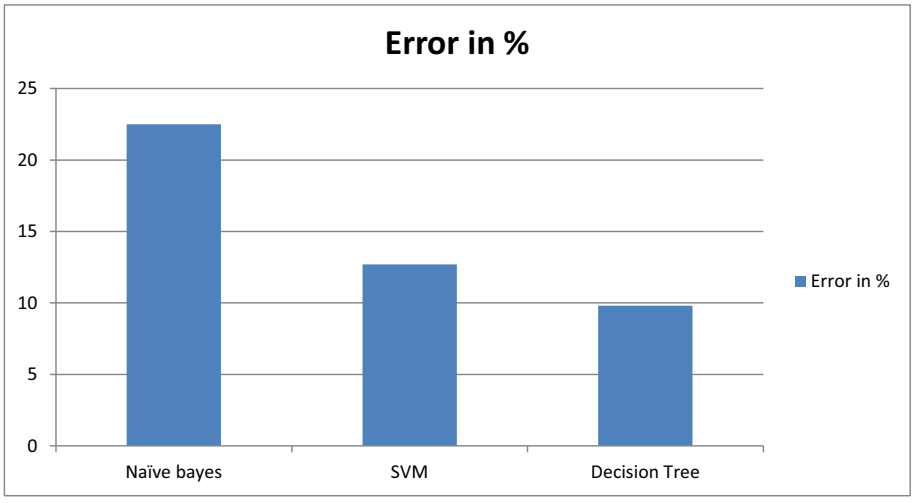

Fig. 2. Error rate results of classification algorithms.

and the leaf node. The produced tree can be used for rule-making. Decision trees are rules that are easy to understand. Many decision-tab algorithms are available, including ID3, C4.5, and CART. The algorithm C4.5 for data mining is a complex decision tree approach. The idea is based on the profit ratio. The main benefits of the C4.5 algorithm are its well-functioning with both categorical and continuous features. It can also handle missing values correctly while running and utilizes less memories. It has the inconveniences of branches that are excessive and insignificant. The information gain is the basis of the ID3 algorithm. CART is a generator of a binary decision tree, which is based on the measure of the Gini index. The ID3 algorithm has discreet features that do not manage missing values.

In the framework, the Cleveland data set [17] is utilized as input. This Cleveland data collection has been preprocessed to eliminate noise and make the data consistent. After preprocessing, the input data is clean and consistent. This data is now fed into machine learning algorithms such as SVM, Nave Bayes, and Decision Tree C4.5. These algorithms classify the data that is sent into them. The classification data is then used as training data for the prediction job. When new patient data is introduced into this framework, the framework predicts whether the new patient's data is normal or abnormal based on the learning data accessible in the classes. It also provides names for possible diseases. Figs. 1 and 2 exhibit the accuracy and error rate of machine learning algorithms.

4. Conclusion

Data mining for healthcare is an interdisciplinary topic of research that evolved from database statistics and is valuable in assessing the efficacy of medical interventions. Data visualization with machine learning Diabetes-related heart disease is a kind of heart disease that occurs in diabetics. Diabetes is a chronic disease that arises when the pancreas fails to create enough insulin or when the body fails to utilize the insulin that is generated appropriately. Heart disease, often known as cardiovascular disease, is a group of disorders affecting the heart or blood arteries. Despite the existence of many data mining classification methods for predicting heart disease, there is insufficient data to predict heart disease in a diabetic individual. We fine-tuned the decision tree model for optimum performance in forecasting the chance of heart disease in diabetic patients since it consistently outperformed the naive Bayes and support vector machine models.

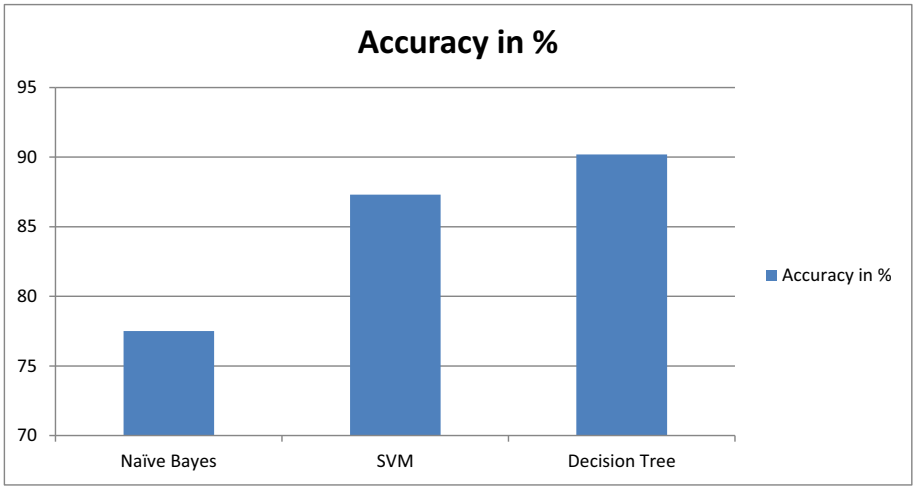

Fig. 1. Accuracy results of classification algorithms.

CRediT authorship contribution statement

K. Arumugam: Visualization. **Mohd Naved:** Data curation. **Priyanka P. Shinde:** Writing - review & editing. **Orlando Leiva-Chauca:** Conceptualization, Methodology. **Antonio Huaman-Osorio:** Writing - original draft. **Tatiana Gonzales-Yanac:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] R. Manne, S.C. Kantheti, Application of artificial intelligence in healthcare: chances and challenges, Curr. J. Appl. Sci. Technol. 40 (6) (2021) 78–89, https://doi.org/10.9734/cjast/2021/v40i631320.
[2] M. Sivakami, P. Prabhu. Classification of algorithms supported factual knowledge recovery from cardiac data set, Int. J. Curr. Res. Rev. 13(6) 161-166. ISSN: 2231-2196 (Print) ISSN: 0975-5241 (Online).
[3] M. Sivakami, P. Prabhu. A Comparative Review of Recent Data Mining Techniques for Prediction of Cardiovascular Disease from Electronic Health Records. In: Hemanth D., Shakya S., Baig Z. (eds) Intelligent Data Communication Technologies and Internet of Things. ICICI 2019. Lecture Notes on Data Engineering and Communications Technologies, vol 38. Springer, Cham 477-484. ISSN 2367-4512 ISSN 2367-4520 (electronic), ISBN 978-3-030-34079-7 ISBN 978-3-030-34080-3 (eBook) 2020.
[4] P. Prabhu, S. Selvabharathi. Deep Belief Neural Network Model for Prediction of Diabetes Mellitus. In 2019 3rd International Conference on Imaging, Signal Processing and Communication, ICISPC 2019 (pp. 138–142) Institute of Electrical and Electronics Engineers Inc. ISBN:9781728136639. 2019.
[5] N. Jothi, N.A. Rashid, W. Husain, Data mining in healthcare – A review, Procedia Comput. Sci. 72 (2015) 306–313.
[6] H. Polat, H. Danaei Mehr, A. Cetin. Diagnosis of chronic kidney disease based on support vector machine by feature selection methods, J. Med. Syst. 41(4) 2017 55.
[7] K.B. Wagholikar, V. Sundararajan, A.W. Deshpande, Modeling paradigms for medical diagnostic decision support: a survey and future directions, J. Med. Syst. 36 (5) (2012) 3029–3049.
[8] E. Gürbüz, E. Kılıç, A new adaptive support vector machine for diagnosis of diseases, Expert Syst. 31 (5) (2014) 389–397.
[9] M. Seera, C.P. Lim, A hybrid intelligent system for medical data classification, Expert Syst. Appl. 41 (5) (2014) 2239–2249.
[10] Y. Kazemi, S.A. Mirroshandel, A novel method for predicting kidney stone type using ensemble learning, Artif. Intell. Med. 84 (2018) 117–126.
[11] H. Barakat, P. Andrew, Bradley, H. Mohammed Nabil Barakat, Intelligible support vector machines for diagnosis of diabetes mellitus, IEEE Trans. Inf. Technol. Bio Med. J. 14 (4) (2009) 1–7.
[12] R. Tina Patil, S.S. Sherekar, Performance analysis of Naive bayes and J48 classification algorithm for data classification, Int. J. Comput. Sci. Appl. 6 (2) (2013) 256–261.
[13] Shruti Ratnakar, K. Rajeswari, Rose Jacob, Prediction of heart disease using genetic algorithm for selection of optimal reduced set of attributes, Int. J. Adv. Comput. Eng. Netw. 1 (2) (2013) 51–55.
[14] S. Grampurohit, C. Sagarnal, Disease prediction using machine learning algorithms, 2020 Int. Conf. Emerg. Technol. (INCET) (2020) 1–7, https://doi.org/10.1109/INCET49848.2020.9154130.
[15] R.J.P. Princy, S. Parthasarathy, P.S. Hency Jose, A. Raj Lakshminarayanan, S. Jeganathan, Prediction of Cardiac Disease using Supervised Machine Learning Algorithms, in: 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 570–575, https://doi.org/10.1109/ICICCS48265.2020.9121169.
[16] P. Deepika, S. Sasikala. Enhanced Model for Prediction and Classification of Cardiovascular Disease using Decision Tree with Particle Swarm Optimization, 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1068-1072, doi: 10.1109/ICECA49313.2020.9297398.
[17] https://archive.ics.uci.edu/ml/datasets/heart+disease.