Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 11, Issue. 2 : 2020
ISSN : **1906-9685**

PREDICTING ILLNESSES USING MACHINE LEARNING ALGORITHMS

[#1]Ms.RAVULA HARITHA, *Assistant Professor*
[#2]Dr.NALLA SRINIVAS, *Assistant Professor*
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,
SREE CHAITANYA INSTITUTE OF TECHNOLOGICAL SCIENCES, KARIMNAGAR, TS.

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 con- dition 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 dia- betic 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.

1. INTRODUCTION

Data collection and processing are especially challenging in the healthcare industry. With the advent of the digital age and technological innovation, a massive amount of patient data in multiple dimensions has been generated. This section contains a wide range of information, including but not limited to patient records, medical history, hospital facilities, and diagnostic data. It is required to handle and assess big, compact, and intricate data in order to get relevant insights that can inform and improve decision-making processes. Discovering hidden patterns in medical data sets is a promising field for medical data mining research.

The application of various data mining techniques and machine learning methodologies has changed healthcare organizations by enabling the discovery of significant patterns and the detection of correlations and interactions among many variables inside enormous databases. The instrument is important in the medical business since it provides and compares existing data for the establishment of future action plans. The technology's integration of several analytical methodologies and cutting-edge algorithms enables the examination of enormous datasets. In the healthcare industry, systematic data collection, organization, and analysis of patient data is used. This strategy can be used to identify the root causes of service delivery issues and the best ways to address them. This could eventually lead to improved diagnostic tools, more precise medical procedures, and more effective treatments. Furthermore, it gives a good foundation for understanding the underlying mechanisms of many elements of the medical industry. With the use of information retrieved from medical databases, disease epidemics can be diagnosed and contained sooner.

Diagnosis is the process of determining what is wrong with a patient based on their indications and symptoms. A diagnostic procedure, or set of diagnostic processes, may or may not include a diagnostic examination. Chronic disease diagnosis is critical in the medical field since it necessitates the study of a wide variety

of symptoms. It's a tough process that frequently leads to false assumptions. Clinical judgment is strongly reliant on both the patient's stated symptoms and the doctors' prior experience and knowledge in disease diagnosis. Furthermore, when new medical systems and therapies are launched, it is becoming increasingly difficult for doctors and other medical professionals to keep up with the rapid rate of change in clinical practice. Medical professionals and doctors must be well-versed in a variety of diagnostic criteria, patient medical histories, and pharmacological therapies in order to provide successful care. However, errors can occur when people disregard formal training and instead rely on gut reactions and assumptions based on limited data and experience with patients. Individuals' cognitive abilities are hampered by a variety of issues, including their inability to multitask, a lack of analytical skills, and short-term memory. As a result, without information from clinical testing and the patient's medical history, it is difficult for a clinician to consistently make the accurate diagnosis. Even the most experienced doctors can benefit from the use of a computer-aided diagnostic system in making more educated and objective diagnoses. As a result, there is a great deal of interest in determining how to outsource the diagnostic process by integrating machine learning techniques with clinicians' experience. Researchers are currently using data mining and machine learning technologies to effectively transform easily accessible data into relevant information in order to improve diagnosis efficiency. Many empirical research have been undertaken to investigate the diagnostic capacities of machine learning systems. Machine learning algorithms have been shown to have a higher diagnostic accuracy rate of 91.1% when compared to highly qualified clinicians (79.97%). Applying explicit machine learning algorithms to datasets relevant to specific diseases yields the best outcomes in disease diagnosis, prediction, prevention, and therapy.

2. RELATED WORK

Bayesian classifiers make predictions by combining a structural model with a set of conditional probabilities. It is assumed that each component's contribution is self-contained. After estimating the prior probability for each class, the process first incorporates the occurrence of each variable value into a previously unknown context. The Bayes network classifier is based on a Bayesian network, which is a graphical model for expressing the joint probability distribution of a set of category attributes.

The SVM algorithm and the Naive Bayes technique were used to predict renal illness. The researchers attempted to characterize the various kidney disease presentations using an algorithm known as an Adaptive Neuro-Fuzzy Inference System (ANFIS). The primary goal of this study was to create a powerful classification system by combining different parameters, such as precision and speed. The Nave Bayes technique outperformed the SVM method in terms of speed, but the SVM method provided more correct classifications. When it comes to forecasting renal illness, the Support Vector Machine (SVM) algorithm surpasses the Naive Bayes technique, according to the findings.

To forecast cardiac disease, a fuzzy technique based on a membership function was applied. The authors employed the Fuzzy KNN Classifier to deal with the existence of ambiguity and uncertainty in the data. The dataset contained 550 records, therefore it was divided into 25 groups, each with 22 items. The dataset was divided along the center to ease training and evaluation. We employed a fuzzy K-nearest neighbors (KNN) technique after a number of preprocessing procedures. This method's efficiency was assessed utilizing a number of evaluation criteria, including accuracy, precision, and recall. The data allowed us to conclude that the fuzzy KNN classifier outperformed the KNN classifier.

Predicting heart illness necessitates a unique technique that employs the Artificial Neural Network (ANN) algorithm. The researchers created an interactive prediction technique based on categorization using an artificial neural network algorithm and thirteen clinically relevant criteria. The proposed methodology has demonstrated its worth as a useful tool for healthcare practitioners in the prediction of cardiovascular illness, with an amazing accuracy rate of 80%.

The researchers used an algorithmic technique to answer complex questions about the

prognosis of cardiovascular illness. This intelligent system's efficacy, accuracy, and desirability were improved using the Naive Bayes technique. This technology may aid doctors in making better educated diagnosis and treatment plans for people suffering from myocardial infarctions. Integration of SMS capabilities, mobile app development for the Android and iOS platforms, and the addition of a pacemaker to the order are all options for upgrading this system.

The adaptability of support vector machines (SVMs) was exploited for diabetic and breast cancer diagnosis. The purpose of adopting adaptive Support Vector Machines (SVM) was to provide an efficient, automated, and flexible diagnostic technique. Modifying the bias value employed by conventional Support Vector Machines (SVM) enhanced the outcomes. As an output, the proposed classifier generates 'if-then' rules. The approach under consideration attained a 100% categorization rate for both diabetes and breast cancer when utilized to diagnose both illnesses. Future research should concentrate on improved ways for controlling the bias parameter in conventional Support Vector Machines (SVM).

Type 2 diabetes incidence can be predicted using a novel method that employs a hybrid model that incorporates clustering and classification approaches. A model combining the K-means clustering algorithm, the C4.5 classification method, and k-fold cross-validation is used to make predictions. The model's results were promising, with a hybrid method yielding a classification accuracy of 88.38 percent. This research has the potential to significantly improve diabetes care by assisting clinicians in making better informed clinical decisions.

3. FRAMEWORK FOR MULTIPLE DISEASE PREDICTION

This framework employs the decision tree, naive Bayes, and support vector machine machine learning approaches.

Bayesian classification is a frequently used probabilistic classification strategy that is based on the application of Bayes' theorem under the premise of independence, however it is not widely known or understood. There is no relationship between the presence or absence of one feature within a category and the presence or absence of any other feature within that category. The system only works in certain situations. Using Bayes' theorem, the probability of one event is computed given the occurrence of another. In the Bayes theorem, let B represent the dependent event and A represent the event that came before it. The value of Sample B in respect to Sample A can be found by dividing the values of both samples by two. This approach of calculating the conditional probability of B given A alone entails dividing the values of both samples by themselves and then dividing by themselves again. The Naive Bayes Classifier requires a small number of training samples to accurately estimate the parameters, in this case the variable medians and variances. Class variances must be calculated due to the assumption of independence. This concept can be used to any number of classes, not simply two or three.

Support Vector Machines (SVMs) are extensively employed in machine learning for kernel learning, an approach for dealing with large-scale prediction issues. The SVM classifier outperformed other classifiers in terms of generalization and scalability for linear and nonlinear data. Furthermore, when paired with a number of well-established methodologies derived from statistical learning and optimization theory, the support vector machine (SVM) classifier performs well in the field of pattern recognition. The purpose of this study is to present a comprehensive perspective that clearly distinguishes between good and bad data, with a focus on the most common error-causing situations. The SVM classification system is based on profit maximization.

When the data is linearly separable, selecting the best hyperplane to divide it into two groups is a piece of cake. Nonlinear mapping via 'Kernel Functions' is what allows Support Vector Machines (SVM) to handle intractable problems by projecting data into a higher dimensional space. Among the various kernel functions accessible are linear kernel functions (LKFs), polynomial kernel functions (PKFs), sigmoid kernel functions (SKFs), and exponential radial basis kernel functions (ERBKFs) (GRBKFs). The Radial Basic Function (RBF) is widely recognized as the greatest kernel function available today.

Decision trees have been widely used in the

classification of large datasets. Classification trees, also known as decision trees, organize data using a tree-like structure, with the "root" node expressing an initial hypothesis and the "leaf" nodes offering "leaf" classification results.

The error rates of the categorization approach were tracked and recorded.

The resulting tree can be used to generate rules. The principles underlying decision trees are simple. There are decision-tree algorithms available, such as ID3, C4.5, and CART. The C4.5 data mining technology is best described by complex decision tree methodology. It is based on a financial return comparison. The C4.5 algorithm's exceptional success is due in large part to its ability to analyze both categorical and continuous data. It consumes less memory than equivalent techniques and gracefully handles missing values while running. Many seemingly insignificant offshoots are regarded as a nuisance. The ID3 algorithm is based on the notion of information The Cleveland dataset is employed as a data source in this strategy. To reduce superfluous details and ensure accuracy, preprocessing methods were applied in the Cleveland data compilation. Following the preprocessing phase, the supplied data is consistent and neat. Support Vector Machines (SVM), Naive Bayes, and the Decision Tree C4.5 are just a few of the machine learning approaches that are currently being used with the available data. These algorithms can classify the data that is fed to them. The classification findings are then utilized to train a model for the prediction job. When a new patient's data is introduced, this framework uses the available learning data inside the classes to produce predictions about the normality or abnormality of the data. It also presents prospective disease labels. Figures 1 and 2 depict the accuracy and error rate of machine learning.

CONCLUSION

Integrating database statistics resulted in the interdisciplinary field of healthcare data mining's creation. It's an excellent resource for determining which therapies are most effective. The use of machine learning algorithms to the problem of data visualization. Diabetes-related cardiovascular disease is referred to as "diabetes-associated cardiovascular illness." acquisition. Computers employ the Classification and Regression Tree (CART) technique to generate a binary decision tree. This decision tree is constructed using the Gini index as a metric to identify which nodes are the most impure. When analyzing discrete features, the ID3 algorithm ignores missing values.

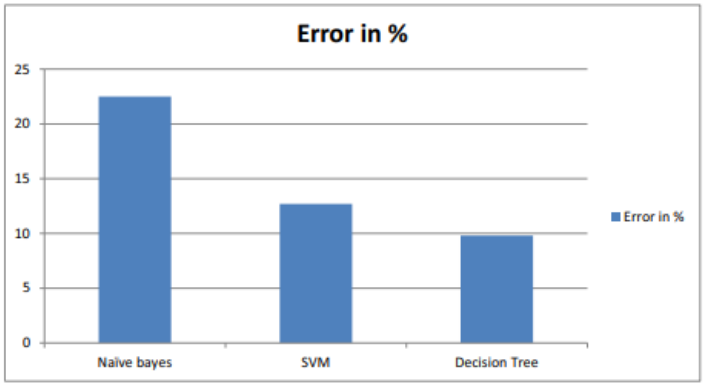

Fig. 2. Error rate results of classification algorithms.

Diabetes is a chronic medical disorder characterized by insufficient insulin synthesis by the pancreas or inefficient utilization of insulin supplied by the body. "Cardiovascular illness," also known as "heart disease," refers to a range of diseases and disorders that affect the cardiovascular system. Although there are several data mining classification approaches for predicting cardiovascular sickness, there is currently insufficient information to create reliable predictions for cardiovascular disease in diabetics. The decision tree model's performance was optimized since it regularly outperformed the naive Bayes and support vector machine models in forecasting the likelihood of heart disease in diabetic individuals.

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.

Fig. 1. The accuracy of classification results.

.

Declaration of Potential Conflicts of Interest

The authors state that they have no financial or other affiliations that could influence the findings presented in this study.