IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 24, Issue 6, Ser. I (Nov. – Dec. 2022), PP 65-69
www.iosrjournals.org

Disease Prediction of Healthcare Data Using Machine Learning Models

Vivekanandan G E
Research Scholar, Computer Science, Engineering, Monad University (U.P) India

Dr Amit Singhal
Research Guide, Computer Science, Engineering Monad University (U.P.) India

ABSTRACT
The challenges faced by medical practitioners in the study of symptoms and early sickness detection stem from the substantial amount of data that has to be processed. Supervised machine learning (ML) algorithms have exhibited significant potential in surpassing traditional approaches for illness detection and aiding healthcare practitioners in promptly detecting high-risk conditions. The primary aim of this research is to discern trends in the diagnosis of illnesses through the use of several supervised machine learning models. This will be achieved by assessing performance metrics associated with these models. The supervised machine learning algorithms that have garnered considerable interest in scholarly discussions encompass Naïve Bayes (NB), Decision Trees (DT), and K-Nearest Neighbor (KNN).
Key Word: *Disease Prediction, Healthcare*

I. INTRODUCTION

The process of evaluating vast volumes of data in order to automatically discover interesting patterns or links is known as "data mining." This, in turn, leads to an increase in the amount of knowledge about the processes that were being examined in the beginning. The practise of data mining may be divided into two distinct categories: Data mining may refer to either the process of mining descriptive data or mining predictive data. The general characteristics of the data that are contained in the database may be generalised or summarised with the help of data mining when a descriptive method is used. The process of mining data for predictions by looking for inferences based on previously collected information is referred to as predictive data mining. The term "data mining" refers to a broad range of operations, some of which include associative rules, classification, forecasting, and clustering, respectively. The process of organising data into categories that have previously been decided is referred to as classification. Classification is a kind of supervised learning. It is one of the most important methodologies in the field of data mining, and it is used extensively in the development of models that estimate future data patterns. To put it another way, this tactic is among the most effective approaches used in data mining. The major goal of the classification techniques is to carry out an analysis on the input data and to generate estimates about the accuracy of the following work. When it comes to the field of medicine, data mining is a procedure that plays an incredibly essential role in establishing the extent to which the patient data and medical data sets included in a massive database are connected to one another. In this study, we classify the liver data set that we obtained from the UCI machine learning repository by doing a comparison analysis of three separate classification techniques: fuzzy logic, decision tree, and fuzzy neural network. These classification strategies are: fuzzy logic, decision tree, and fuzzy neural network. Fuzzy logic, fuzzy decision trees, and fuzzy neural networks are some of the categorization approaches that fall under this category.

When it comes to the gathering and processing of patient data, the healthcare business is one of the ones that faces the greatest amount of difficulty. As a consequence of the dawn of the digital age and the steady advance of scientific discovery, there has been an uptick in the production of a large quantity of data with several dimensions pertaining to individual patients. The clinical features, hospital resources, information on the diagnosis of illness, patients' records, and medical equipment are all included in this data. Before any meaningful knowledge can be retrieved from the vast quantity of information that is both dense and intricate, it is necessary to process and analyse the information. Only then can the knowledge be used in decision making. The discovery of previously unknown patterns in medical data sets is a substantial potential that may be investigated via medical data mining.

LEARNING UNDER SUPERVISION

The research of opinions may be seen as a more orderly activity, in which case machine learning methods are a potentially useful computational tool for addressing this matter. One subcategory of machine learning methodologies examines pre-existing data to discover how individuals express their emotions via language. After then, it leads to the establishment of the borders between the different groupings. It incorporates a model that may be used to analyse concealed data in order to characterize such data in accordance with the viewpoint that they convey. According to statements made by Elkan and Charles in 2010, the objective of a supervised learning algorithm is to acquire a classifier via the process of learning from a set of training instances.

Predictions based on test instances may be generated using a classifier if one is available. The supervised approach is the name given to this particular technique. In this method, the data is provided in the form of a labelled dataset, from which a model may be trained to generate the result of the problem in question. The use of supervised learning allows for the problem space to be partitioned into two distinct categories: regression problems and classification problems. The value of a distinguishing variable may be anticipated with the use of a categorization code. It is feasible to think of the input data as being a part of a certain class or group. Take, for example, a collection of photographs of animals, each of which is identified as a tiger, lion, etc. In this particular scenario, the algorithm has to place the newly uploaded photographs into one of these categories. The following are some examples of classification algorithms:

- Logistic-Regression
- Naïve-Bayes-Classifier
- Support-Vector-Machines
- Logistic-Regression-Problems

The regression procedure is used while working with continuous data. Estimating the cost of renting an apartment in a certain city by taking into account factors such as the neighborhood, the size, the kind of building, and so on. The data that has been gathered is then sent to the computer, which makes a price prediction for the flat based on models that have been used in the past.

The analysis of patient data in healthcare is an extremely important component in the process of improving patient outcomes and promoting evidence-based decision making. Using data from healthcare systems, this research study investigates how machine learning models might be used to perform categorization and illness prediction tasks. The purpose of the project is to create accurate models that are able to categorise patients in an efficient manner and efficiently forecast the consequences of illness based on a variety of patient characteristics, medical records, and diagnostic information. The technique for this study includes preprocessing and analysing a large healthcare dataset, training and testing several machine learning algorithms, and evaluating the performance of these algorithms using acceptable evaluation measures. The findings of this research have the potential to improve the decision-making process in the healthcare industry, provide support to clinical research, and bring about improvements in patient care. In recent years, the healthcare sector has seen a huge development in both the amount and complexity of healthcare data. This growth has been one of the most notable trends. These records include a broad variety of patient information, including as demographics, medical history, laboratory findings, diagnostic imaging, and treatment records. The process of gleaning relevant insights from such enormous volumes of data may be an intimidating one for individuals working in the healthcare industry.

CLASSIFICATION AND DISEASE PREDICTION IN HEALTHCARE

A Definition and goals of classification and disease prediction tasks:

- The process of giving predetermined labels or categories to new occurrences depending on the characteristics of those instances is known as classification. Classification models are useful in the medical field for a variety of purposes, including the prediction of the existence of a certain ailment as well as the categorization of patients into various disease classifications.
- The purpose of disease prediction is to arrive at an estimate of the chance that a person will get a certain illness within a specified amount of time. Machine learning models may generate personalized risk scores or probabilities by analyzing patient features and risk variables. This can be accomplished by analyzing the data.

B. Potential benefits in healthcare decision-making and patient care:

- The process of medical practitioners making more informed decisions on the care and treatment of individual patients may be aided by classification and sickness prediction models, which may give support to the practitioners.

- Early disease diagnosis, which makes it feasible to take quick preventive and therapeutic steps, is made possible by the application of prediction models. This, in turn, leads to improved patient outcomes and reduced total healthcare expenses.
- These models might assist with resource allocation by identifying high-risk patients who would need more intensive monitoring or intervention. In other words, they could help find people who need more support. Using this information, one may choose who should get more focus and consideration.
- With the aid of prediction models, medical professionals are able to provide customised therapy and tailored treatments because they are able to stratify patients according to the risk profiles that are associated with them.
- These models have the ability to improve diagnostic accuracy while simultaneously reducing the number of diagnostic errors that occur as a result of their exact identification of the existence of a disease or its progression.
- Triaging patients, which helps determine which patients need immediate treatment and optimises the delivery of medical care, may benefit from classification models. This helps patients get the care they need in the most timely manner.
- Disease prediction models contribute to the management of population health by identifying groups of individuals who are at risk and giving information that might potentially impact preventative efforts taken by public health agencies.

Classification and sickness prediction models have the potential to significantly impact both the decision-making process in the healthcare industry and the course of therapy that patients receive. They make early detection feasible, as well as customized treatments, the best allocation of resources, and better diagnostic accuracy. customized therapies are also made possible by these technologies. By using these models, medical personnel are able to improve the results for their patients, speed up the operations of the healthcare system, and enhance their capacity to monitor the health of a whole community.

OBJECTIVES

1. To investigate the potential of machine learning models in improving healthcare decision-making and patient outcomes.
2. To identify important features and patterns in healthcare data that contribute to accurate classification and disease prediction

II. RESEARCH METHODOLOGY

A model that has been proposed for the purpose of predicting preterm birth, often known as PTB. investigate the concept that features that are part of a feature subset that is being investigated cannot include characteristics that have a meaningful relationship between them. In this study, it was also said that the information gain that the traits have is related to the degree to which they may be regarded independent from one another. This was discussed in relation to the study's main topic, which was information gain. It is predicted that as a consequence, it will yield superior results when compared to data that has not been examined. This is the expectation. The obstetrical dataset, which refers to information associated with the process of giving birth, is the major focus of the current inquiry.

DATA ANALYSIS

The assessment of several classifiers, namely Logistic Regression (LR), Decision Trees (DT), and Support Vector Machines (SVM), utilised in this study is performed by evaluating their specificity, accuracy, and sensitivity [24, 158]. The evaluation of the suggested model may be conducted by comparing it to the three designated classifiers using the stated performance metrics. The performance metrics of many classifiers for both the unaltered dataset and the equilibrated dataset are displayed in Table 1 and Table 2 correspondingly.

Table 1 Performance Metrics of the Classifiers-Original Dataset

classifier	accuracy	Sensitivity	specificity
DT	0.777	0.702	0.930
LR	0.841	0.863	0.971
SVM	0.861	0.801	0.702

Table 2 Performance Metrics of the Classifiers- Balanced Dataset

classifier	accuracy	Sensitivity	specificity
DT	0.796	0.713	0.972
LR	0.872	0.832	0.954
SVM	0.909	0.891	0.783

Based on the data shown in Table 5 and Table 6, it can be deduced that the three learning classifiers exhibit an approximate accuracy of 85%. Based on an examination of the primary dataset, it is apparent that the support vector machine (SVM) has the highest level of accuracy, reaching 86.1%. This surpasses the performance of both logistic regression (LR) and decision tree (DT) techniques.

The utilisation of the Synthetic Minority Over-sampling Technique (SMOTE) on a dataset that is balanced, containing both term and preterm samples, has the potential to further improve the outcomes, as mentioned in reference. The performance of Support Vector Machines (SVM) demonstrates a significant enhancement, increasing from 86.1% to 90.9%, when assessed on a balanced dataset in contrast to the initial dataset. Thus, in summary, it is evident that the Support Vector Machine (SVM) model emerges as the most efficient classifier employed in this experimental study. was to develop a prediction model for Preterm Birth (PTB) in pregnant women through the use of machine learning methodologies. The suggested model may be utilised to identify the noteworthy maternal traits (linked to preterm birth) present in the obstetrical dataset by the implementation of an entropy-based feature selection approach. In order to achieve this objective, three learning classifiers, specifically Support Vector Machine, Logistic Regression, and Decision Tree, are utilised to categorise all instances of birth into two groups: Term Birth (TB) and Preterm Birth (PTB). Upon evaluating the accuracy of the three aforementioned classifiers, it becomes apparent that the Support Vector Machine (SVM) demonstrates the highest degree of competence, with an accuracy rate of 90.9%.

The main aim of the proposed model is to offer support to healthcare professionals in making well-informed decisions throughout the delivery of maternity care. The procedure entails the detection of risk variables linked to preterm delivery and delivering prompt notification to pregnant individuals. The use of a proactive strategy has been shown to be beneficial in mitigating many possible challenges that may arise during pregnancy. This method not only helps in minimising the costs associated with diagnostic procedures but also significantly reduces the probability of Preterm Birth (PTB). A noteworthy constraint of this study is the relatively modest scale of the obstetrical dataset, which encompasses a restricted number of significant factors linked to preterm delivery. The aforementioned limitation should be considered as a promising route for future study and investigation.

III. CONCLUSION

The development of Medical Disease Diagnosis Systems (MDDSs) has been pursued to address categorization challenges in healthcare data by applying several machine learning approaches. However, the task of selecting appropriate learning algorithms for the purpose of analysing healthcare data is a substantial problem in the field of data mining applications. Medical datasets are currently being analysed using data mining approaches to find significant patterns that have the potential to provide useful information. Subsequently, these patterns are utilised for the purpose of clinical data diagnosis. Clinicians have a significant hurdle when attempting to make a diagnosis simply based on the current medical condition of a patient, particularly in the absence of comparative cases. The available empirical evidence provides robust support for the proposition that medical data sets are intrinsically situated within the domain of the natural sciences. The data sets under consideration exhibit notable characteristics like their substantial volume and disparate distribution, along with the existence of disputes, ambiguity, and significant intricacy. Therefore, the task of developing an appropriate and all-encompassing predictive model for disease diagnosis utilising these datasets presents considerable difficulties. Therefore, the main objective of this dissertation is to develop a capable Medical Disease Diagnosis System (MDDS) that can successfully address the aforementioned challenges.

REFERENCES

[1]. Yadav, S. S., & Jadhav, S. M. (2020). Detection of common risk factors for diagnosis of cardiac arrhythmia using machine learning algorithm. Expert Systems with Applications, 163, 113807.
[2]. Masood, A., Sheng, B., Li, P., Hou, X., Wei, X., Qin, J., & Feng, D. (2018). Computer- assisted decision support system in pulmonary cancer detection and stage classification on CT images. Journal of biomedical informatics, 79, 117-128.
[3]. Yadav, D. C., & Pal, S. (2020). Prediction of thyroid disease using decision tree ensemble method. Human-Intelligent Systems Integration, 1-7.
[4]. Liu, T., Fan, W., & Wu, C. (2019). A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. Artificial Intelligence in Medicine, 101, 101723.
[5]. Fang, G., Liu, W., & Wang, L. (2020). A machine learning approach to select features important to stroke prognosis. Computational Biology and Chemistry, 88, 107316.
[6]. 45. Chen, Y. C., Suzuki, T., Suzuki, M., Takao, H., Murayama, Y., & Ohwada, H. (2019). Building a classifier of onset stroke prediction using random tree algorithm. International Journal of Machine Learning and Computing, 7(4), 61-66.
[7]. Govindarajan, P., Soundarapandian, R. K., Gandomi, A. H., Patan, R., Jayaraman, P., & Manikandan, R. (2020). Classification of stroke disease using machine learning algorithms. Neural Computing and Applications, 32(3), 817-828.
[8]. Arslan, A. K., Colak, C., & Sarihan, M. E. (2019). Different medical data mining approaches based prediction of ischemic stroke. Computer methods and programs in biomedicine, 130, 87-92.
[9]. Goyal, M. (2018, July). Long short-term memory recurrent neural network for stroke prediction. In International Conference on Machine Learning and Data Mining in Pattern Recognition (pp. 312-323). Springer, Cham.

[10]. Li, X., Wu, M., Sun, C., Zhao, Z., Wang, F., Zheng, X., ... & Zou, J. (2020). Using machine learning to predict stroke- associated pneumonia in Chinese acute ischaemic stroke patients. European Journal of Neurology, 27(8), 1656-1663.
[11]. Heo, J., Yoon, J. G., Park, H., Kim, Y. D., Nam, H. S., & Heo, J. H. (2019). Machine learning–based model for prediction of outcomes in acute stroke. Stroke, 50(5), 1263- 1265.
[12]. Liu, Y., Yin, B., & Cong, Y. (2020). The Probability of Ischaemic Stroke Prediction with a Multi-Neural-Network Model. Sensors, 20(17), 4995.
[13]. Monteiro, M., Fonseca, A. C., Freitas, A. T., e Melo, T. P., Francisco, A. P., Ferro, J. M., & Oliveira, A. L. (2018). Using machine learning to improve the prediction of functional outcome in ischemic stroke patients. IEEE/ACM transactions on computational biology and bioinformatics, 15(6), 1953-1959.
[14]. Shah, T., Yavari, A., Mitra, K., Saguna, S., Jayaraman, P. P., Rabhi, F., & Ranjan, R. (2019). Remote health care cyber-physical system: quality of service (QoS) challenges and opportunities. IET Cyber-Physical Systems: Theory & Applications, 1(1), 40-48.
[15]. Nandyala, C. S., & Kim, H. K. (2019). From cloud to Fog and IoT-based real-time U- healthcare monitoring for smart homes and hospitals. International Journal of Smart Home, 10(2), 187-196.