IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 1, 2022

Machine Learning for Early Detection and Prevention of Chronic Diseases

Afaque Alam[1] ,Dr. Savya Sachi[2]

[1]Assistant Professor, Department of CSE
Bakhtiyarpur College of Engineering, Bakhtiyarpur, Bihar, India
Email id- afaque.alam.cse@gmail.com
[2]Assistant Professor, Department of CSE
Princeton Institute of Technology, Narapally, Hydrabad, Telangana, India

Abstract- In the field of biomedical and healthcare communities, accurate prediction plays a major role to find out the patient's risk of the disease; the only way to overcome the mortality due to chronic diseases is to predict it earlier so that the disease prevention can be done; such a model is a patient's need in which machine learning is highly recommended; however, the precise prediction on the basis of symptoms becomes too difficult for doctor. Technological development, including machine learning, has a huge impact on health through an effective analysis of various chronic diseases for more accurate diagnosis and successful treatment. The hardest difficulty is accurately predicting sickness. Data mining is crucial in predicting the sickness in order to solve this issue. Using a chronic illnesses dataset from the UCI machine learning data warehouse, this work applies machine learning techniques to the analysis of chronic diseases. We employ data mining approaches to construct dependable prediction models for chronic illnesses, such as diabetes, cancer, heart disease, and kidney disease.

Keywords— *Logistic Regression, Chronic Diseases, Machine Learning, Diseases Prediction, Accuracy.*

INTRODUCTION

Programming computers to operate optimally using sample or historical data is known as machine learning. Studying computer systems that learn from data and experience is known as machine learning. Machine learning is divided into two categories: supervised learning, which involves producing output variables based on input variable prediction, and unsupervised learning, which involves grouping distinct groups together for a specific intervention. By identifying intricate models and extracting medical information, machine learning (ML) introduces experts and specialists to new concepts. ML prediction models have the potential to enhance guidelines for making decisions about the care of individual patients in clinical settings. They can also independently diagnose various illnesses in accordance with clinical guidelines. By integrating these models into medication prescriptions, physicians can save time and benefit from new identification opportunities in medicine. It has been demonstrated that machine learning is a useful tool for assisting in the analysis and forecasting of the vast amounts of data generated by the healthcare sector. We optimize machine learning methods to anticipate chronic illness outbreaks more accurately. A few publications only provide a cursory overview of using ML approaches to predict sickness. We provide a unique approach that seeks to improve the accuracy of illness prediction by utilizing machine learning techniques including K-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Logistic Regression, Random Forest, and Naive Bayes (NB) to identify key characteristics. To increase the learning process' accuracy, several of these algorithms are run. Then, using the accessible datasets, it may be tested. When the prediction model is presented, various combinations of characteristics and a range of established categorization methods. With ML models, it can also be possible to improve quality of medical data, reduce variation in patient rates, and save in medical costs. Therefore, these models are frequently used to investigate diagnostic analysis when compared with other conventional methods. To reduce the death rates caused by chronic diseases (CDs), early detection and effective therapy are the only solutions. Therefore, most medical scientists are attracted to the new technologies of predictive models in disease estimation.

These new advancements in medical care have been spreading the accessibility of electronic data and opening new doors for decision support and productivity improvements. ML methods have been effectively utilized in the computerized elucidation of pneumonic capacity tests for the differential analysis of CDs. It is expected that the models with the highest accuracies could gain large importance in medical diagnosis.

LITERATURE SURVEY

The name machine learning was coined in 1959 by Arthur Samuel. Tom Mitchell states machine learning as " Machine learning is the study of computer algorithms that allow computer programs to automatically improve through experience" . It is a combination of correlations and relationships. Most machine learning algorithms in existence are concerned with finding and/or exploiting relationship between datasets. Once Machine Learning Algorithms can pinpoint on certain correlations, the model can either use these relationships to predict future observations or generalize the data to reveal interesting patterns. There are various types of algorithms such as Linear Regression, Logistic Regression, Naive Bayes Classifier, KNN (K-Nearest Neighbor Classifier), Decision Tress, Entropy, SVM (Support Vector Machines), K-means Algorithm, Random Forest etc. Machine learning examine the study and construction of algorithms that can learn from and make predictions on data. It is closely related to (and often overlaps with) computational statistics, which also focuses on prediction-making through the use of computers. It has strong ties with mathematical optimization which delivers methods, theory and application domains to the field. Machine learning is sometimes merged with data mining, where the latter subfield focuses more on exploratory data analysis and is known as unsupervised learning. The study for the best medical diagnosis mining technique was performed by K.M. Al-Aidaroos, A.A. Bakar, and Z. Othman. For this study, the authors compared Nave Baeyes to five other classifiers: LR, KStar (K*), Decision Tree (DT), Neural Network (NN), and a basic rule-based algorithm (ZeroR). The efficiency of all algorithms was evaluated using 15 real-world medical problems from the UCI machine learning repository (Asuncion and Newman, 2007). In the experiment, NB outperformed the other algorithms in 8 of the 15 data sets, leading to the conclusion that the predictive accuracy results in Nave Baeyes are superior to other techniques. Darcy A. Davis, Nitesh V. Chawla, Nicholas Blumm, Nicholas Christakis, and Albert-Laszlo Barabasi discovered that treating chronic illness at a global level is neither time nor cost effective. As a result, the authors performed this study in order to forecast potential disease risk. CARE (which uses only a patient's medical history and ICD-9-CM codes to predict possible disease risks) was used for this. Based on their own medical history and that of similar patients, CARE incorporates collective filtering approaches with clustering to predict each patient's greatest disease risks. ICARE, an iterative version that integrates ensemble principles for improved efficiency, has also been defined by the authors. These cutting-edge systems don't need any advanced knowledge and can predict a wide range of medical conditions in a single run. ICARE's remarkable potential risk coverage means more precise early alerts for thousands of illnesses, several years ahead of time. When used to its full extent, the CARE system can be used to investigate a wider range of disease backgrounds, raise previously unconsidered questions, and facilitate discussions regarding early detection and prevention. This research paper was written by JyotiSoni, Ujma Ansari, Dipesh Sharma, and SunitaSoni to provide a survey of existing techniques of information discovery in databases using data mining techniques that are used in today's medical research, specifically in Heart Disease Prediction. A number of experiments have been carried out to compare the performance of predictive data mining techniques on the same dataset, and the results show that Decision Tree outperforms, with Bayesian classification having comparable accuracy to Decision Tree in some cases, but other predictive approaches such as KNN, Neural Networks, and Classification based on Clustering underperform. Shadab Adam Pattekari and Asma Parveen conducted a study to predict heart diseases using the Decision Tree Algorithm, in which the consumer provides data that is compared to a qualified set of values. As a result of this study, patients were able to provide basic information that was compared to data, and heart disease was expected. M.A.NisharaBanu and B. Gomathy analysed the various types of heart-related problems using medical data mining techniques such

ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 1, 2022

as association rule mining, grouping, and clustering I. The aim of a decision tree is to show any possible outcome of a decision. To achieve the best result, various rules are devised. The criteria used in this study were age, sex, smoking, being overweight, drinking alcohol, blood sugar, heart rate, and blood pressure. The risk level for various parameters is saved with their ids ranging from 1 to 100. (1-8). The standard level of prediction is represented by IDs less than 1, whereas higher IDs other than 1 represent higher risk levels. The pattern in the dataset is studied using the K- means clustering method. The algorithm divides the data into k groups. The closed cluster is allocated to each point in the dataset. Each cluster centre is recalculated as the average of the cluster's points.

PROPOSED SYSTEM

For this study, we have used structured and unstructured data from the healthcare domains to estimate the risk of illness. re-creating missing data in medical records from internet sources using a latent component model. We might also use statistical data to evaluate the most common chronic illnesses in a certain region and demographic. To find out about helpful aspects for working with structured data, we speak with hospital specialists. We employ the random forest technique to automatically choose features in unstructured text files.

(i) Data collection- To diagnose the sickness, data has been gathered from the internet; here, the true symptoms of the illness that is, no dummy values are gathered. The disease's symptoms are gathered from several websites that deal with health.

(ii) Building Model- Data mining is done using a variety of techniques. Machine learning is among the methods. Among the various machine learning techniques used in random forests are summarization, grouping, and clustering. categorization is one of the data mining procedures in this phase of categorical data categorization since classification techniques are employed in this project. Additionally, there are two stages to this step: testing and training. Predefined data and related class labels are utilized for categorization during the training phase.

(iii) Prediction- Prediction done by Random Forest Model using Flask frame work model trained by training chronic disease dataset

METHODOLOGY

This work presents a precise system for the detection of CKD through the utilization of a robust model. The proposed approach leverages ML techniques to construct a prediction model that is both effective and accurate. To visually depict the various stages of the proposed system, Figure 1 provides a schematic representation.

(i) Data Collection- In order to validate our proposed ML model, we obtained the CKD dataset from the UCI ML Repository. The dataset contains a total of 400 samples, which we used for evaluating and validating our ML model in this study. Each sample comprises 24 predictive variables, including 11 numerical variables and 13 categorical (nominal) variables. The dataset also includes a categorical response variable called ‘ class’ , which indicates the presence or absence of CKD. The ‘ class’ variable has two distinct values: ‘ ckd’ for samples diagnosed with CKD and ‘ notckd’ for samples without CKD.

(ii) Preprocessing- Medical datasets are prone to various issues that can have a negative impact on the performance of ML models. Therefore, it is crucial to address these challenges to improve the quality of the data. The preprocessing stage plays a vital role in enhancing data quality by tackling key issues such as data encoding, missing values, and outliers.

(iii) Data Encoding- To handle the combination of categorical and numeric features in the dataset, the label encoder module from the Scikit-learn library was used. This module transformed the categorical features into numeric representations, allowing for the improved performance of the machine learning model.

ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 1, 2022

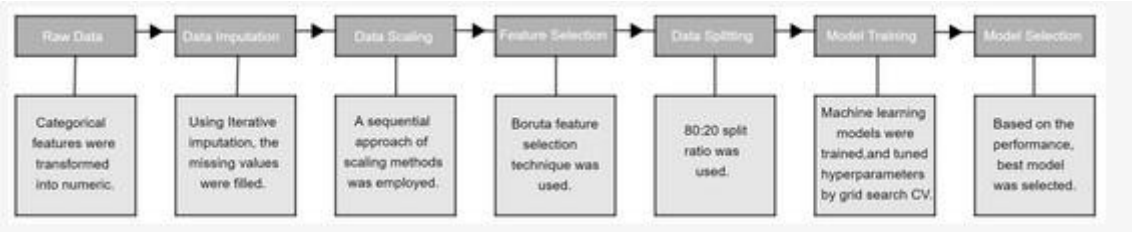

Figure 1- Proposed workflow

Algorithm 1 The iterative imputation pseudocode.

Input:

1: Dataset X

2: Features with missing values: F_{missing}

3: Maximum iterations: η

4: Convergence threshold: ϵ

Output: Imputed dataset X_{imputed}

5: **procedure** IterativeImputation(X, F_{missing}, η, ϵ)

6: Initialize $X_{\text{imputed}} \leftarrow X$

7: **for** each feature f in F_{missing} **do**

8: Initialize missing mask M_f for feature f

9: Initialize model M_f (Linear Regression) for feature f

10: Initialize convergence ←False

11: Initialize iterations ←0

12: **while** not convergence and iterations < η **do**

13: Fit model M_f on X_{imputed}

14: Predict missing values using M_f

15: Update X_{imputed} with predicted values

16: Check for convergence using mean absolute change

17 : **if** CheckConvergence(X_{imputed}, f, ϵ) **then**

18: convergence ← True

19: **end if**

20: Increment iterations

21: **end while**

22: **end for**

23: **return** X_{imputed}

24: **end procedure**

CONCLUSION

The healthcare industry has seen significant advancements because to machine learning. The diagnosis of chronic diseases is one of the challenging and vital activities that machine learning helps to make simple and dependable. It has resulted in groundbreaking modifications to laboratory, clinic, and hospital protocols. Through previous and current data analysis, physicians may forecast their patients' future circumstances. We have evaluated our technique on many datasets related to diabetes, kidney disease, heart disease, and cancer. This study's primary goal was to use features to predict chronic illness while keeping a greater accuracy (in this case, an accuracy of almost 90%). Additionally, our algorithm produces a report with the likelihood of a

ISSN PRINT 2319 1775 Online 2320 7876
Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 1, 2022

condition occurring. The outcomes show how reliable the suggested method is. To improve the prediction of chronic diseases, future research should examine several supervised and unsupervised machine learning techniques with additional performance criteria.

REFERENCES

[1] Hamet P., Tremblay J. Artificial intelligence in medicine. Metabolism. 2017; 69:S36– S40. doi: 10.1016/j.metabol.2017.01.011.

[2] Johnson K.W., Soto J.T., Glicksberg B.S., Shameer K., Miotto R., Ali M., Dudley J.T. Artificial intelligence in cardiology. J. Am. Coll. Cardiol. 2018; 71:2668– 2679. doi: 10.1016/j.jacc.2018.03.521.

[3] Bini S. Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? J. Arthroplast. 2018; 33:2358– 2361. doi: 10.1016/j.arth.2018.02.067.

[4] Kotsiantis S.B., Zaharakis I., Pintelas P. Supervised machine learning: A review of classification techniques. Emerg. Artif. Intell. Appl. Comput. Eng. 2007; 160:3– 24.

[5] Deo R.C. Machine Learning in Medicine. Circulation. 2015; 132:1920– 1930. doi: 10.1161/CIRCULATIONAHA.115.001593.

[6] Battineni G., Sagaro G.G., Nalini C., Amenta F., Tayebati S.K. Comparative Machine-Learning Approach: A Follow-Up Study on Type 2 Diabetes Predictions by Cross-Validation Methods. Machines. 2019; 7:74. doi: 10.3390/machines7040074.

[7] Polat H., Mehr H.D., Cetin A. Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods. J. Med. Syst. 2017; 41:55. doi: 10.1007/s10916-017-0703-x.

[8] A.Davis, D., V.Chawla, N., Blumm, N., Christakis, N., & Barbasi, A. L. (2008). Predicting Individual Disease Risk Based On Medical History.

[9] Adam, S., & Parveen, A. (2012). Prediction System For Heart Disease Using Naive Bayes.

[10] Al-Aidaroos, K., Bakar, A., & Othman, Z. (2012). Medical Data Classification With Naive Bayes Approach. Information Technology Journal.

[11] Darcy A. Davis, N. V.-L. (2008). Predicting Individual Disease Risk Based On Medical History.

[12] JyotiSoni, Ansari, U., Sharma, D., & Soni, S. (2011). Predictive Data Mining for Medical Diagnosis: An Overview Of Heart Disease Prediction.