IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁙ Impact Factor 8.102 ⁙ Peer-reviewed / Refereed journal ⁙ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124139

Chronic Disease Prediction using Machine Learning

Chakrapani D S[1], Kruthika M Hiremath[2], Megana N[3], Nandini H T[4], Nanda D C[5]

Asst.Professor, Computer Science & Engineering, Jawaharlal Nehru New College of Engineering, Shivamogga, India[1]

Computer Science & Engineering, Jawaharlal Nehru New College of Engineering, Shivamogga, India[2-5]

Abstract: One of the biggest issues facing the healthcare industry is chronic diseases. The global populace consumes a lot of unhealthy food. Generally speaking, doctors must thoroughly review the patient's records in order to diagnose the ailment. Sometimes it is difficult for doctors to treat patients effectively since the diagnosis is manual. Chronic disease patients are becoming more and more numerous every day. Therefore, environmental threat assessment is important. Currently, the digitization of healthcare is taking advantage of advances in medical care in hospitals. The traditional style lacks the knowledge development to monitor and analyze problems and views of traditional situations and has been replaced by the process of gaining better understanding from clinical data through the use of predictive analytics and fact. time machine readable tools. Reduce the size to include conditions like heart, stroke, diabetes, and cancer across multiple datasets using keywords selected in the research using machine learning penalty debt. Feature selection plays an important role in machine learning by selecting key features for live event detection.

Keywords: MachineLearning; sklearn; Randomforest; Linear regression.

INTRODUCTION

Machine learning has proven effective in supporting decision making and forecasting from the large volumes of data generated by the healthcare industry. We simplify machine learning algorithms to effectively predict disease outbreaks. Many studies only give us an idea of the possibility of using machine learning techniques to predict diseases. We present a new method to improve the accuracy of disease prediction using machine learning such as KNN, decision tree (DT), logistic regression, random forest and pure Bayesian (NB) to see the main features. Many of these algorithms are used to improve the accuracy of the study. It can then be tested with existing data. Prediction models with turnover, and save medical costs. Therefore, this method is frequently used in clinical research compared to other methods. Early diagnosis and effective treatment are the only solutions to reduce the death rate from chronic disease (CD). Therefore, most medical researchers are interested in new predictive model methods in disease prediction. These new developments in healthcare have expanded accessibility to electronic information and opened new doors for decision support and increased productivity. The ML method has been successfully applied in computer interpretation of lung examinations for different CDs. The most accurate samples should take priority in diagnosis.

With advancing technology and analytical capabilities, medical digitization provides new insights to create a variety of tools and resources to improve health through specialized software such as Electronic Health Records (EHR). A lot of medical information from Big data can help medical doctors, including patient information, medical history, medical records and demographics. broken. In this article, many materials have been reviewed for the classification of chronic diseases. Although chronic disease is usually controlled by medical treatment, it cannot be cured due to its long duration. Chronic diseases such as diabetes mellitus, stroke, kidney disease, cancer, heart disease and arthritis are leading causes of death as health problems often need to be managed. Early diagnosis is necessary to detect new problems associated with the disease through appropriate medical management and good care. These records are stored in large electronic medical records containing records of different traits or variables. Various special options are used for various classification problems. These features have a high value when using distribution for estimation because this can be avoided by reducing features using multiple selection methods This approximation will also help identify important diseases related to for good prognosis. Feature selection is a method originally used in machine learning to help the class better predict disease. The prediction and classification of the disease was trained and tested in samples using different machine learning classification algorithms with optimized feature selection to achieve high accuracy. optimization is the task of selecting the best target function from a set of tasks by reducing or optimizing the function and comparing it with the different options available. to the system. By optimizing the parameters in the transmission, a search algorithm system finds the best solution and achieves good results. In this review article, algorithms .Feature selection plays an important role in pretraining the model to make predictions. Various options are available in three categories: filters, wraps, and embedding techniques. Filtering methods in the selection feature reveal the latest trends in search based on rankings

DOI: 10.17148/IJARCCE.2023.124139

without relying on machine learning algorithms. The method is fast and can achieve better results with fewer large files. The wrapper method uses the training model to generate a score for a set of features. The performance of the wrapper method for every new feature subset is evaluated by the machine learning algorithm. This method will shows better accuracy in performance by finding out the best features in detecting their dependencies. The embedded method benefits the approach of filter and wrapper as the feature subset are chosen based on the learning algorithm to train the model by deriving the important features for prediction.

machine learning algorithms are used to predict disease duration across different datasets. The dataset used in the current model is preprocessed, features are selected as important and used in model for classification. Classification algorithms such as Support Vector Machines, Logistic Regression, Random Forests are used to determine whether a patient has the disease Results are evaluated according to accuracy and time.

LITERATURE REVIEW

This section reviews the research literature in describing chronic diseases using machine learning algorithms to solve predictive problems. Logistic regression examines the relationship between independent variables and a dependent variable and evaluates the probability of the relationship by fitting the data to a single variable. The system takes medical advice from a person who can come from blood tests. In addition, for example, whether a person smokes regularly, drinks alcohol, etc. Some personal questions need to be answered. Based on such ideas, the output can be predicted whether the person has cancer or not.

(SushmithaManikandan.,2017) [4] Heart Attack prediction system. Rapid Miner is used to process files first. Rapid Miner is a data science software platform that provides an integrated platform for predictive analytics, text mining, deep learning, and machine learning. Gaussian Naive Bayes algorithm from the scikit-learn package was used for

(I.Preethi,Dr.K.Dharmarajan..,2020)[1] Diagnosis of classification. Scikit-learn covers many data mining tools.

(Imesh Udar Ekanayake, Damayanthi Herath.,2020) [2] Chronic kidney Disease prediction using machine learning methods. The missing values should be based on their distribution to achieve the required accuracy. In this study, the K nearest neighbor imputer algorithm is used to fill in the missing values. When using the algorithm, the initial distribution of the data is stored using the same number of estimates as the number of all events, given the smallest and smallest standard deviation. In this study, 11 classification models in education are discussed. The data is divided into 3 parts as 70% training data, 15% cross validation data and 15% test data. The algorithm with the highest accuracy was selected in all 3 data sets. This study introduces new functionality, including prior knowledge, no processing fees, and specific options for positively or negatively predicting CKD. In addition, this study highlights the importance of integrating information into the selection process when analyzing clinical data on kidney disease.

(Shraddha S. More, Vivian Brain lobo, Ronald Melwin laban Simran Panchal,Monika patil, Gulshan Pathak..,2020) [3] Cancer prediction and insurance eligibility using machine learning. Machine learning is a part of AI that mainly focuses on developing logical processes. In this study, logistic regression, k-NN and DT algorithms are used. These algorithms are used to process current data and generate historical data for output. The system developed for cancer prediction is trained to read data and use input from users. The results obtained are whether the patient has cancer or not. KNN is used for both regression and classification Categorizes data according to probability and similar properties. It uses points to find common points and splits these points into packages. DT algorithms can be used to solve both regression and classification problems.

User Interface is the Web interface is used to obtain the risk factor of the patient. An interactive web interface was developed to access the classifier and predict the risk factor of the individual. The data follows a normal distribution, Gaussian Naïve Bayes algorithm was used for the classification. Amongst all the state of the art classification algorithms applied on the dataset, Naïve Bayes has proven to give the best accuracy of 81.25%. The same classifier can be accessed with the help of a web interface for convenience of the user. As part of the future of this system, the new system proposes that classification algorithms can be used to improve accuracy.

(Priyanka Sonar, prof. KjayaMalini.,2019) [5] Diabetes prediction using different machine learning approach. The classification algorithms are decision trees, support vector machines, pure Bayesian and neural networks for correct classification. The training data in Machine Learning is used to train the model for rich processing. Get detailed information from the training process to train the model. Therefore, these models are combined into prototypes. Preprocessing refers to changes made before feeding our data into an algorithm. Preprocessing techniques are used to transform raw data into understandable. Feature Extraction is used to convert input data into output of features. The characteristic square measure is a feature of design ideas and helps to differentiate classes of design ideas. The target database is the updated database instead. For example, you set up a certificate Upgrade Source database named demo.

DOI: 10.17148/IJARCCE.2023.124139

Next, you get a copy of the production file. Then copy the content changes from the demo database to the copy. The demo database here is your source and the target is the replica.

(Vijeta Sharma, Shrinkala Yadav, Manjari Gupta.,2020) [6]

User Interface is the Web interface is used to obtain the risk factor of the patient. An interactive web interface was developed to access the classifier and predict the risk factor of the individual. The data follows a normal distribution, Gaussian Naïve Bayes algorithm was used for the classification. Amongst all the state of the art classification algorithms applied on the dataset, Naïve Bayes has proven to give the best accuracy of 81.25%. The same classifier can be accessed with the help of a web interface for convenience of the user. As part of the future of this system, the new system proposes that classification algorithms can be used to improve accuracy.

(Priyanka Sonar, prof. KjayaMalini.,2019) [5] Diabetes prediction using different machine learning approach. The classification algorithms are decision trees, support vector machines, pure Bayesian and neural networks for correct classification. The training data in Machine Learning is used to train the model for rich processing. Get detailed information from the training process to train the model. Therefore, these models are combined into prototypes. Preprocessing refers to changes made before feeding our data into an algorithm. Preprocessing techniques are used to transform raw data into understandable. Feature Extraction is used to convert input data into output of features. The characteristic square measure is a feature of design ideas and helps to differentiate classes of design ideas. The target database is the updated database instead. For example, you set up a certificate Upgrade Source database named demo. Next, you get a copy of the production file. Then copy the content changes from the demo database to the copy. The demo database here is your source and the target is the replica.

(Vijeta Sharma, Shrinkala Yadav, Manjari Gupta.,2020) [6]

Heart disease prediction using machine learning techniques The purpose of the article is to develop a machine learning system to recognize heart disease based on risk factors. We used the UCI Heart Disease Prediction comparison dataset in this study, which includes 14 different types of heart disease. Machine learning algorithms such as random forests, support vector machines (SVM), naive Bayes and decision trees were used for modeling. Four popular machine learning methods chosen for generating heart disease prediction models are: Support Vector Machine, Decision Trees, Naive Bayes, Random Forest Classification. In the data collection step, the Cleveland Cardiology dataset available online in the UCI repository was used in the research study

PROPOSED SYSTEM

chronic diseases progress slowly, it is important to predict early and give the right medicine. Therefore, it is necessary to propose a decision-making model that can help identify chronic diseases and predict future patient outcomes. Although there are many approaches to this problem in the field of AI, this study clearly demonstrates the ML prediction model for long-term diagnosis. We will be able to find useful results with ML algorithms that improve the quality of patient data and the analysis of certain medicinal products compared to traditional data analysis methods. The main goal of our work is to develop software that is efficient and feasible to simplify work in the hospital and replace the forecasting process with medical management. Our programs help doctors increase efficiency, reduce medical errors and waste time. If the disease is predictable, the patient can be treated early, thus reducing the risk of life and saving the patient's life. With early diagnosis, the cost of treatment can also be reduced to some extent.

We perform diagnostics based on various classification machine learning models such as:

- Linear regression
- Random Forest

Linear regression is a method for estimating the value of one variable against the value of another variable (predictive analytical method). Linear regression makes predictions of continuous/real or numerical variables. The linear regression algorithm shows the relationship between one variable (y) and one or more variables (y), hence it is called linear regression. Since linear regression shows a relationship, it means that the variable changes according to the value of the variable.

Linear regression models provide slopes that represent the relationship between variables. Consider the following figure. Linear Regression in Machine Learning We can represent linear regression as:

$$y = a0 + a1x + \varepsilon$$

where Y= sum of variables, a1 = linear regression coefficient (ratio of each value.) ε = random error The values of x and y variables are training datasets for linear regression model representation.

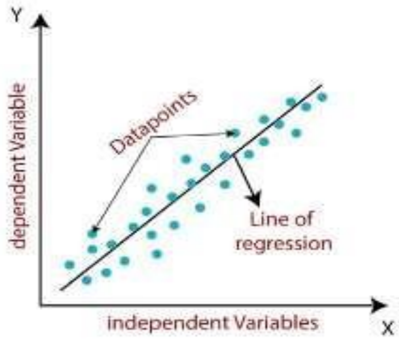


fig 1. Linear Regression

The random forest algorithm is based on the prediction of the decision tree. It is based on the concept of ensemble learning, which is the process of combining multiple classifiers to solve complex problems and improve model performance. A random forest is a classifier that has many decision trees in various subsets of a given dataset and averages them to improve the prediction accuracy of the dataset. The following diagram explains how the Random Forest algorithm works

Fig 2. Random Forest

METHODOLOGY

The architecture diagram is a visual representation of the actual physical appearance of a software system component. This diagram describes the system software in terms of system monitoring.

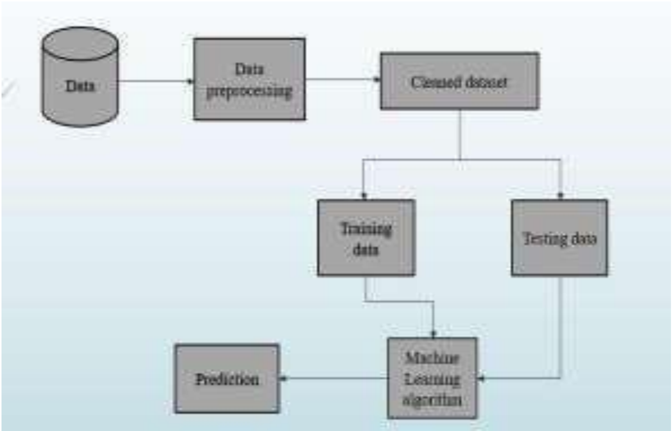


Fig 3. System framework

DOI: 10.17148/IJARCCE.2023.124139

Data: Data is an essential part of machine learning. It refers to a set of observations or measurements that can be used to train a machine learning model. The quality and quantity of data available for training and testing plays an important role in determining the performance of machine learning models. The data can use a variety of data such as numerical, categorical or time series data and can come from a variety of sources such as databases, spreadsheets or APIs. Machine learning algorithms use data to learn patterns and relationships between different inputs and outputs that can be used for prediction or classification.

Data preprocessing: Data prioritization is the process of organizing raw data and fitting it into machine learning models. This is the first important step in designing a machine learning model. We don't need to examine clean and formatted data when creating machine learning. Before doing anything with the file, it should be cleaned and put in a formatted form. We use previous data for this.

Data Cleaning: Data cleaning is the process of preparing data for examination by removing or replacing inaccurate, incomplete, irrelevant, duplicate or incorrect information. However, as we said above, it is not as simple as adjusting some lines or deleting files to make room for new files. Keeping records is a lot of work. Stop guessing without water, there's a reason data cleaning is the most important step if you want to build a data culture. It includes:

- Correcting spelling and grammatical errors
- Creating a standard document
- Fixing errors such as blank spaces

- Checking for duplicate documents or some mixed patterns counted by the bidding method, your information may contain text or comments.

Test data: Training data is data used to fit the model. The validation certificate is data used to measure how well the model fits the data affected by hyperparameters. Evaluation has increased as the ability to analyze data is an important part of standard setting.

Machine Learning Algorithms: A machine learning algorithm is how an intelligent machine usually works by predicting the output value given the input data. Prediction: In machine learning, prediction referred as an output of an algorithm that learns from historical data. The algorithm then generates values for the unknown variable in each new data set.

Feature selection: The next step in developing predictive models for chronic diseases is architecture. Feature selection is the process of selecting relevant features from data and converting them into a format that machine learning algorithms can use. Trait selection involves selecting the traits most important for disease benefit and removing redundant or redundant traits. Feature selection also includes normalizing or normalizing data to ensure that all features are in a similar state.

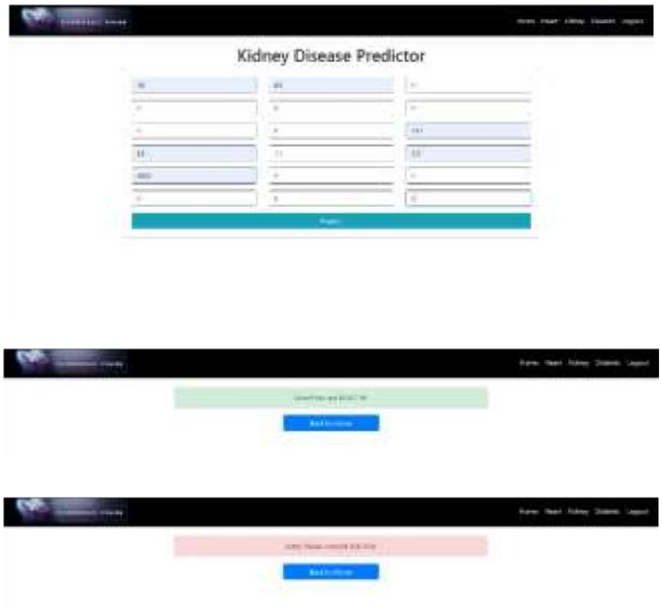
Model selection: After choosing a feature, the next step is to choose an appropriate machine learning model. The choice of machine learning model depends on the type of data, the size of the data, and the probability of disease. Many machine learning models for chronic disease prediction include logistic regression, decision trees, random forests, and neural networks. Metrics such as accuracy, precision, recall and F1 score can be used to compare the performance of different models.

Model Training and Validation: When choosing a machine learning model, it needs to be trained on the data. The dataset is generally divided into training and testing processes, the training process is used to train the model, and the testing process is used to evaluate the model's performance. The hyper parameters of the model are tuned using techniques such as cross validation to improve its performance. The performance of the model is evaluated using indicators such as accuracy, precision, recall and F1 score.

RESULTS

The random forest model was found to be the best performing model with 95.25% accuracy. The high accuracy and recall rates indicate that the model is effective in identifying people at risk for chronic diseases and accurately identifying those not at risk

DOI: 10.17148/IJARCCE.2023.124139



Kidney Disease Predictor

Fig 4. Snapshot of results

CONCLUSION

Chronic diseases are the leading cause of death because people are not aware of early diseases. A diagnosis of is required because it cannot be treated with the appropriate drugs. This article examines different selections and classification algorithms of methods for estimating chronic diseases using a large amount of data. Distributions used in the current study to predict chronic diseases include Linear regression and random forest. A prediction engine that allows a user to check if they have diabetes or heart disease. The user interacts with the prediction engine by typing a document that holds the parameter set provided as input to specify the model. Prediction engine provides the best performance compared to other state-of-the-art methods. The prediction engine predicts the presence of diseases using three algorithms: Linear Regression, Random Forest.

REFERENCES

[1] Dr..K Dharmarajan, I. Preethi, "Diagnosis of chronic disease in a predictive model using machine Learning algorithm", International conference on smart Technologies in computing ICSTCEE 2020, IEEE 2020.

[2] Imesh Udar Ekanayake, Damayanthi Herath, "Chronic kidney Disease prediction using machine learning methods", moratesva engineering Research conference 2020, IEEE 2020.

[3] Shraddha S. More, Vivian Brain lobo, Ronald Melwin laban ,Simran Panchal,Monika patil, Gulshan Pathak. "Cancer prediction and insurance eligibility using machine learning" , Fifth International conference on communication and electronic systems (ICCES 2020) IEEE conference record #48766.

[4] Sushmitha Manikandan,"Heart Attack prediction system",International conference on energy communication, data Analytics and soft computing (ICECDS-2017) IEEE2017.

[5] Priyanka Sonar, prof.K.jayaMalini "Diabetes prediction using different machine learning approaches", third international conference on computing methodologie and communication (ICCMC 2019),IEEE 2019.

[6] Vijeta Sharma, Shrinkala Yadav, Manjari Gupta "Heart disease prediction using machine learning techniques", International Conference on advance in computing communication control and networking (ICACCCN) IEEE 2020.

[7] Njoud Abdullah Almansour, Hajra Fahim syed, Nuha Radwan Khayat, Rawan Kanaan Altheeb "Neural network and support vector machine for the prediction of chronic kidney disease", computers in Biology and medicine 2019.

[8] Gopi Battineni, Getu Gamo Sagaro, Nalini Chinatalapudi and Francesco Amenta"Applications of machine learning Predictive Models in the Chronic Disease Diagnosis", Journal of Personalized Medicine 2020.

[9] Divya Krishnani, Anjali Kumari, Akash Dewangan ,Aditya Singh, Nenavath Srinivas Naik “Prediction of Coronary Heart Disease Using Supervised Machine Learning Algorithms” Conference(TENCON) IEEE 2019.

[10] Divya jain,Vijendra Singh “ Feature selection and classification system for chronic disease prediction ” , Egyptian informatics journal 2018.