WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Breast Cancer Disease Prediction Using Machine Learning Algorithm

Ms. V. Abarna[1], Mrs. J. Jeyachidra[2]

II MCA[1], Department of Computer Science and Applications,
Dean faculty of[2], Department of Computer Science and Applications,
Periyar Maniammai Institute of Science and Technology, Vallam, Thanjavur, Tamil Nadu,India
abarnasona7899@gmail.com, deanscse@pmu.edu

ABSTRACT:

A multidisciplinary topic of study with origins in database statistics, data mining for healthcare can be used to evaluate the efficacy of medical therapies. Many of the currently available machine learning models for healthcare analysis concentrate on one condition at a time. For instance, one analysis might focus on thyroid issues, another on diabetes, and yet another on cancer conditions. There is no approach that can predict numerous diseases with a single analysis. This project offers a Python Flask API-based system for predicting a variety of diseases. This work made use of diabetes, thyroid, and breast cancer analysis. Multiple sickness analysis was carried out using machine learning algorithms, tensor flow, and the Flask API. Python pickling is used to load the pickle file, while Python save model behavior is used to save model behavior. The relevance of this research is that it evaluates disorders and includes all of the parameters that cause the condition, allowing the disease's maximum impact to be detected. We conduct a thorough search of all available feature variables in the KAGGLE dataset to construct models for cardiovascular, prediabetes, and diabetes identification. Using several time-frames and feature sets for the data (based on laboratory data), the Support Vector Machine algorithm is used to forecast diseases with greater accuracy

Keywords—Disease Prediction Machine Learning Algorithm Support vector machine (SVM)

I. Introduction

The Support Vector Machine (SVM) technique is widely used for multiple ailment prediction in the healthcare and medical informatics industries. SVM is a supervised machine learning approach used for regression and classification tasks. To predict the simultaneous presence or absence of various diseases, SVM is employed in multiple disease prediction. This method can help medical professionals diagnose patients more accurately and quickly while also detecting and preventing sickness early on. Finding a hyperplane that splits the data points into the most distinct classes is how the SVM method operates. The algorithm is trained on a dataset that contains details on a patient's presence or absence of various diseases in the case of multiple sickness prediction. other clinical elements, including age, gender, and medical background. Once trained, the SVM model may use clinical data to predict the likelihood of a variety of diseases in a new patient. A probability score is generated by the SVM model for each disease, indicating the likelihood that the patient will have that disease. Compared to conventional diagnostic techniques, multiple illness prediction using the SVM algorithm has a number of benefits, including a reduced risk of misdiagnosis and increased diagnosis accuracy. Additionally, it can help with the creation of individualized treatment plans and the identification of individuals who are at high risk for certain diseases. Overall, this approach has the potential to reduce healthcare expenses while simultaneously considerably improving healthcare outcomes.

II. EXISTING WORK&PROPOSED WORK

Existing work

One of the most important applications of machine learning algorithms today is disease detection and treatment. Machine learning techniques are also utilized to discover connections and links between diseases. Many people are dying today as a result of diabetes and cardiovascular disease. Prediction and diagnosis of diabetic and cardiac disease has become a difficult task for doctors and hospitals both in India and abroad. To reduce the number of deaths caused by diabetes and heart disease, we must first determine whether a person is at risk of developing diabetes and heart disease. Data mining techniques and machine learning algorithms are extremely significant in this field. In this present system, emphasis is placed on how data mining techniques can be applied. The process of seeking to determine and/or identify a suspected disease or disorder, as well as the decision reached by this process, is a key task of any diagnostic system. Machine learning methods are frequently employed for this. To be useful in medical diagnostic problems, these machine learning approaches must have excellent performance, the ability to deal with missing and noisy data, the transparency of diagnostic information, and the ability to explain conclusions. As people generate more data every day, there is a need for a classifier that can reliably and efficiently

classify freshly generated data. This system primarily focuses on the supervised learning technique known as Random forests for data classification by modifying the variables.

Proposed work:

Because machine learning models have the potential for powerful predictive analytics, they open up new avenues for healthcare. We apply supervised machine learning models to predict diabetes, thyroid illness, and cancer disease in this project. Despite the acknowledged link between these diseases, we designed the algorithms to predict cancer, thyroid, and diabetes individually to benefit a broader spectrum of patients. As a result, we can uncover feature similarities between diseases that affect their prediction. Prediction of pre-diabetes and undiagnosed diabetes is also taken into account. In addition, this study investigates a support vector machine model that combines the outcomes of numerous supervised learning models to improve prediction ability. Multiple supervised learning models are used in this study to classify at-risk patients. In supervised learning, the learning algorithm is fed training data that includes both the recorded observations and the labels that correspond to the category of the observations. This information is used by the algorithm to construct a model that, when presented with fresh data, can predict which output label should be linked with each new observation. Support Vector Machines (SVM) classify data by dividing it into classes and separating them with a boundary, such as a line or multi-dimensional hyperplane. Optimisation ensures the greatest possible boundary separation of classes. While SVM frequently beats logistic regression, the model's computational complexity necessitates lengthy training times for model building.

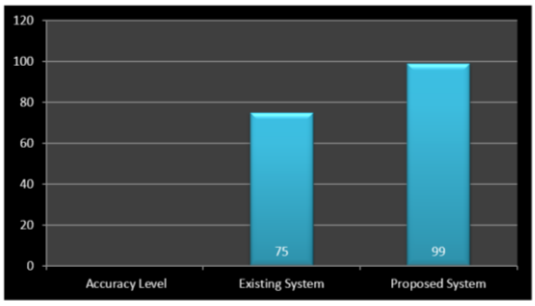

Fig no:1

III. System Architecture

The conceptual model that specifies a system's structure, behavior, and other perspectives is referred to as a system architecture, sometimes known as a systems architecture. A formal description and representation of a system that is organized to make it feasible to reason about its structures and behaviors is called an architecture description. System architecture may contain system elements, their observable characteristics, and the connections (such as interactions' behavior) between them. It can offer a plan for getting things and making systems that work well together to complete the whole system. Initiatives have been made to formalize architectural description languages (ADLs), which are languages for specifying system architecture.

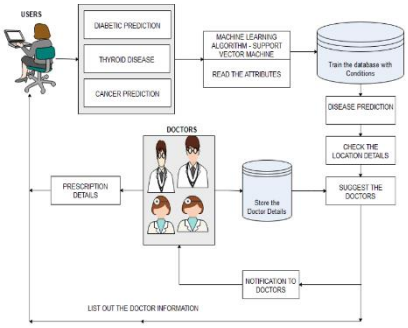

Fig no :2

IV. Methodology

A. Framework Constructions:

Create the admin and user logins with this module. Admin can upload datasets relevant to heart disease and diabetes. A data set (or dataset, as this form is not found in many modern dictionaries such as Merriam-Webster) is a collection of data. A data set is most usually associated with the contents of a single database table or a single statistical data matrix, where each column of the table represents a specific variable and each row refers to a specific

member of the data set in question. Values for each variable, such as an object's height and weight, are present for every component of the data collection. A datum is a name for any value. The data set may include information for one or more members, depending on the number of rows. The phrase data set can also refer to the data in a group of closely related tables referring to a specific experiment or occurrence. We can add datasets relevant to diabetic, thyroid, and cancer disorders in this module, which comprise attributes like age, gender, height, weight, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoking, alcohol, active status, and cardiac labels.

B. Data Pre-processing:

Data pre-processing is an important step in the [data mining] procedure. The adage "garbage in, garbage out" is especially relevant to initiatives involving data mining and machine learning. The methods used to collect data are frequently not tightly regulated, which leads to out-of-range numbers, impossible data combinations, missing information, etc. Data analysis that has not been thoroughly checked for these issues may yield false results. Therefore, before performing an analysis, it is crucial to consider the representation and quality of the data. If there is a lot of redundant, unimportant information accessible, or if the data is noisy, knowledge discovery during the training phase is more difficult. Data preparation and filtering tasks can take a long time to complete. In this module, we can estimate the missing values of and remove the irrelevant variables. Offer structured datasets at the end.

C. Feature Selection:

The process of finding the most significant inputs or of limiting inputs for processing and analysis is known as feature selection. The practice of removing relevant information or features from existing data is referred to as feature engineering (or feature extraction) in a related phrase. The features are sorted according to the score and either kept in the dataset or deleted using filter feature selection methods, which utilize statistical measures to assign each feature a score.

D. Classification

In this module, a classification method is used to predict diabetic and heart diseases, and a machine learning technique, such as the Support vector machine algorithm, is used to forecast the diseases. A SVM is a feed forward vector model that maps input data sets to appropriate outputs. It is made up of numerous vectors of nodes in a directed graph, and each layer is completely connected to the one before it. The value of each feature represents the value of each data item, which is plotted as a point in n-dimensional space (where n is the number of characteristics you have). a specific position in the SVM algorithm. Then, classification is achieved by identifying the hyper-plane that most effectively separates the two classes. Support vectors are computed using the coordinates of each individual observation. A frontier that effectively separates the two classes (hyper-plane/line) is the SVM classifier. The user can supply the features and the system will automatically anticipate the diseases.

E. Recommendation:

Every year, the healthcare business creates gigabytes of data. The medical documents kept are a repository of information about patients. The process of obtaining meaningful information or providing quality treatment is difficult and critical. At the moment, regular body diagnostic is required to maintain wellness. There are numerous sources accessible today as standalone prediction or recommendation systems, but the need of the hour is to have an integrated model that incorporates both. Furthermore, rather than visiting hospitals and clinics on a regular basis, it would be more appropriate and handier if consumers could acquire basic diagnoses online 24 hours a day, seven days a week. As a result, costs are reduced and time is saved. If specific irregularities are discovered in the diagnosis, recommendations of nearby specialists and hospitals based on the user's preferences would aid in the timely and right treatment. Healthcare, as a subject that is constantly evolving and producing a great amount of data, creates a need to use the data for meaningful knowledge, attracting large organizations to invest heavily in this field. The Support Vector Machine algorithm uses these symptoms to forecast disease. Furthermore, all necessary and sufficient information about the expected sickness, as well as the recommended doctors, is presented. Recommendation provides the location, contact information, and other relevant details of illness specialists based on the filters selected by the user from less fees, more experience, and the doctors' nearby location.

Algorithms used

As the name implies, we use the Support Vector Machine (SVM) for classification and the Multilinear Regression (MLR) for prediction in our disease prediction system. MLR is a type of regression method in which multiple independent values are used to predict a value based on two or more factors.

A single Independent/Predictor(X) variable is utilized to model the response variable (Y) in simple linear regression. However, there may be many occasions where the reply variable is affected by many forecaster variables; in such scenarios, the MLR algorithm is used.

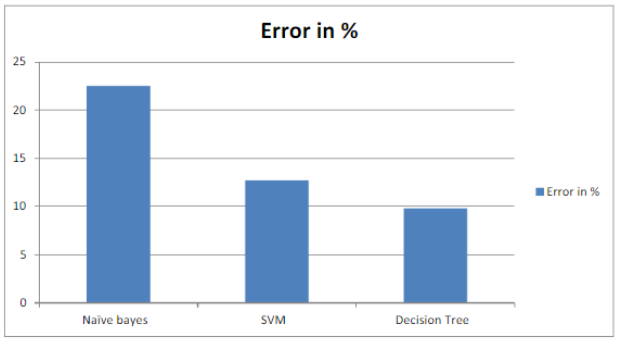

Fig no:3

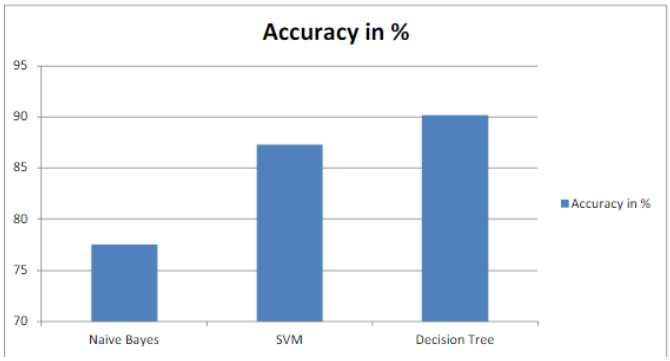

Fig no:4

Conclusion

Data mining in medical data analysis is a great approach to look at existing relationships between variables. We've demonstrated that mining can help us discover relevant connections even when the features we're looking for aren't direct indicators of the class we're attempting to forecast. In our study, we attempted to estimate the chance of establishing a system for forecasting diabetic, thyroid, and cancer disease datasets, and we show that the suggested method improves disease prediction accuracy. This type of classifier can help in the early detection and prediction of diabetic patients. Patients can be warned to modify their lifestyle in this manner. This will result in the prevention of many diseases, leading in lower mortality rates and lower health-care costs for the state. SVMs have been demonstrated to be a classification technique with great prediction performance, and they have also been explored using the ROC curve for both training and testing data. As a result, our SVM model can be recommended for disease classification and recommending doctors based on disease forecasts. Extend the framework to include many diseases and recommend diagnosis information such as doctor recommendations, prescriptions, and so on.

Future enhancement

We can develop the framework in the future to add different deep learning algorithms, which will increase accuracy rates. We can also examine the system to anticipate various diseases. Extend the framework to include many diseases and recommend diagnosis information such as doctor recommendations, prescriptions, and so on

Limitations

- Labelled data-based disease classification
- Provide high number of false positive
- Binary classification can be occurred
- Computational complexity

References

R. Katarya and P. Srinivas, “Predicting heart disease at early stages using machine learning: A survey,” in 2020 International Conferenceon Electronics and Sustainable Communication Systems (ICESC), 2020,pp. 302–305

N. Lavesson, Evaluation and Analysis of Supervised Learning Algo-rithms and Classifiers, 2006.

V. S and D. S, “Data Mining Classification Algorithms for Kidney Dis-ease Prediction,” International Journal on Cybernetics & Informatics,vol. 4, no. 4, pp. 13–25, 2015

R. Chen, "Support Vector Machines", Artificial Inteligence, 2022.

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,

] 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. 2019

D. Dahiwade, G. Patle, and E. Meshram, “Designing disease prediction model using machine learning approach,” Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 2019, no. Iccmc, pp. 1211–1215, 2019

D. Yao, J. Yang, and X. Zhan, “A novel method for disease prediction: Hybrid of random forest and multivariate adaptive regression splines,” Journal of Computers (Finland), vol. 8, no. 1, pp. 170–177, 2013.