2023 International Conference on Inventive Computation Technologies (ICICT) | 979-8-3503-9849-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICICT57646.2023.10133957

Cardiovascular Disease Risk Assessment using Machine Learning

Nikkila Prakash
Department of Networking and Communications
SRM Institute of Science and Technology,
Chennai, India
nikkilaprakash.phy@gmail.com

Mohitth Mahesh
Department of Networking and Communications
SRM Institute of Science and Technology
Chennai, India
mohitthmahesh@gmail.com

Dr. Gouthaman.P
Department of Networking and Communications, Faculty of Engineering and Technology,
SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu 603203, India
Chennai, India
gouthamp@srmist.edu.in

Abstract- **Cardiovascular diseases (CVD) are one of the highest causes of death in the world. The early detection of cardiac risk is a critical factor in proper diagnosis and treatment. This way, patients with critical needs get priority access to doctors and healthcare systems. In this study, a cardiac risk assessment system was developed using the Logistic Regression algorithm, a machine learning model that has a high accuracy and is easy to interpret. The dataset used in this study included information from various patients. A total of 13 features were used to train the Logistic Regression model, including age, gender, blood pressure, and cholesterol levels. The results demonstrated that the Logistic Regression algorithm achieved high accuracy in predicting CVD risk, with an accuracy of 86.89. The main challenge when it comes to CVD risk assessment is the complexity of algorithms which makes it difficult for healthcare practitioners to interpret the results. Some systems require the personnel to go through additional training to use the risk assessment system, which can be time consuming. Logistic Regression is straightforward and simple. It is easy to interpret, making it suitable for clinical settings. It also has a well-established framework, which makes it very practical and reliable. This study showcases the importance of machine learning in the field of healthcare and highlights the effectiveness of the Logistic Regression algorithm in predicting cardiac risk. The high accuracy achieved by the model enables the early identification of cardiovascular disease risk. This makes it a useful tool for the healthcare industry and public health initiatives.**

Keywords-Cardiovascular Disease, Logistic Regression, Machine Learning, Risk Assessment

I. INTRODUCTION

One of the biggest causes of death all around the world is cardiovascular disease (CVD). By detecting the presence of disease early and taking action to intervene, one can reduce the fatalities caused by cardiovascular diseases. This can reduce the overall risk of the disease.

Recently, machine learning has been emerging as a technology that can be leveraged to predict outcomes and used as a tool to reduce the risk factor of such diseases. By using machine learning, data of a patient can be analyzed to identify individuals that are at risk. These predictions can be used by healthcare providers to make informed decisions and take preventive steps. By using machine learning, the accuracy of assessing the risk of heart disease can be improved. This can lead to treatment and preventive strategies that can prove to be very effective.

The aim of this study is to make an interface that is open to the public using which they can analyze their risk factor for cardiovascular disease, with suggestions to improve cardiovascular endurance. This can help keep the public informed about their health and encourage preventative action.

I. LITERATURE SURVEY

According to [1], there are a lot of researchers currently using data mining techniques to help with the predictions in healthcare industries. Using a large amount of data ensures that different demographics are studied. Here, the decision tree model has achieved the highest accuracy. They determined that the J48 algorithm is the best classifier for this scenario, after comparing it with the logistic model tree and random forest algorithm. This is important because it shows the potential of using machine learning in predicting heart diseases, and its effectiveness when combined with data mining techniques. One of the problems with this research is that only a small number of decision tree techniques were considered and compared. Other popular algorithms like CART or C5.0 could've been considered.

A study that highlights feature selection is [2], it emphasized the importance of selecting particular features to develop models. In this study, they compare and contrast between different algorithms to compare the accuracies. They have compared naive bayes, random forest, support vector machines, decision trees, k-nearest neighbors and ensemble model. They came to the conclusion that different models perform in different ways according to the scenario in hand. This highlights the importance of selecting an algorithm according to the problem in hand. A drawback in this research is the lack of including ensemble algorithms in their comparison. Some research into using ensemble networks for disease prediction could have yielded better conclusions.

One study [3] uses a decision tree algorithm to predict the probability of disease. They use the map reduce algorithm. In this study, they use both unstructured and structured data. Upon comparing their model with other widely used algorithms like convolutional neural networks, they determined that their system had a higher accuracy. This highlights the importance of comparing the model with others to ensure the highest accuracy can be achieved. One of the disadvantages in this work is that the MapReduce algorithm is good only for batch processing

One of the important areas in the healthcare prediction system is using data from all demographics to ensure accurate prediction. [4] had the objective of using data mining combined with machine learning models to determine the probability of the outcome variable. Here they focus on prediction of heart disease. In this study they focus on building a model with the Random Forest algorithm after data preprocessing. On comparing their model with other widely used algorithms, they concluded that Random Forest had the highest accuracy. The limitation with this research is the lack of implementing more ensemble and complex models. This could've produced higher accuracy than the existing models.

Some studies such as [5] use a hybrid model by combining two machine learning algorithms. In this study the objective was to predict the probability of contracting a particular disease with a high accuracy rate. After preprocessing the data they used a combination of support vector machine algorithm and multilinear regression for the prediction system. Upon comparing the hybrid model with various algorithms such as convolutional neural networks, decision trees and k-nearest neighbors, they determined that the hybrid model displayed a higher accuracy for predicting the likelihood of a disease. One of the demerits of this work is that it was tested using the minimum amount of data possible. The accuracy can be improved by using a larger amount of data which could take a very long time to collect and use.

According to [6], The Latent Factor Model is used to recreate missing data in medical records obtained from online libraries and other sources. The study uses a mixture of structured and unstructured data to determine the risk of disease. The data obtained is pre-processed by using several steps. One such step is checking for null values and filling them using the forward fill method. The prediction itself is performed using a Random Forest Algorithm and a Flask framework. The limitation in this work is that comparative analysis of different models and algorithms weren’t performed. The random forest model was chosen without analysis of other suitable models.

The study performed in [7] uses the symptoms entered as the input and predicts the probability of a disease and provides the same as an output. The study uses the Decision Tree Classifier to predict the probability of disease. The study also highlights the potential of the method in early diagnosis and analysis of medical data. One of the demerits in this research is the zero frequency problem.This occurs when the model assigns zero to a categorical variable if the category wasn’t available in the training dataset.

According to [8], Machine Learning Algorithms are used to predict Parkinson’s Disease using Voice Data. The classification of the dataset is done through the K- Nearest Neighbor and Support Vector Machine Algorithms. The study shows the possibility of early diagnosis of Parkinson’s Disease using voice analysis of patients. The limitation in this work is that some methods of unsupervised learning are not able to produce high accuracy results with the available voice dataset. The availability and accessibility of voice based datasets is also quite less.

Another study performs the comparison of several machine learning algorithms for disease prediction [9]. The study compared algorithms such as Naive Bayes, Random Forest, Support Vector Machine and Logistic Regression. It found that the Random Forest Algorithm and Logistic Regression had the best accuracy out of all the algorithms tested. The study also highlighted the importance of feature selection and extraction in the improvement of performance of these algorithms. The research was only carried out on studies that applied several machine learning models on the same data. This was due to the fact that each study had different research scope and datasets. This limits the scope of the results achieved. This is the drawback of this work.

According to [10], Using hybrid models obtained by combining various machine models like Random Forest and Linear Method improves the accuracy of the model over using other algorithms. They achieve a higher accuracy than K-Nearest Neighbor, Naive Bayes and Decision Trees. They found hybrid models tend to have better performance in terms of feature extraction and feature selection. According to the research, 13 key attributes or clinical features have been used as the input. As a hybrid model is being used, The model could be highly resource intensive and network intensive. This is one of the limitations of this work.

The research and study performed in [11] seeks to describe the different approaches such as machine learning and deep learning and their applications in cardiovascular medicine. One of the concerns about using AI in healthcare is the lack of data privacy, selection bias and poorly selected datasets.

According to [12], Several different theoretical models of deep learning are described in order to better apply the process of deep learning to cardiovascular diseases. Models such as Restricted Boltzmann Machine (RBM) and Auto-encoder are described. One of the limitations as mentioned by the research is the lack of data available which often results in lower accuracy.

The research done in [13] seeks to prove that cardiovascular diseases can be predicted using classic risk factors through data mining and machine learning. This could be preferred instead of time-consuming, invasive or even dangerous medical examinations or tests. The drawback in this research is that the relationship between cardiac risk factors and cardiovascular diseases are not linear and hence more studies are required in order to apply machine learning in healthcare systems.

The work performed in [14] aims to explore data-driven approaches to predict cardiovascular diseases. The National Health and Nutrition Examination Survey (NHANES) dataset was used to search for all the available features in order to develop models for cardiovascular disease prediction

According to [15], Deep Neural Network is used to make predictions for heart diseases. A comparison of different models such as K-nearest neighbor, Random Forest,Support Vector Machine and Logistic Regression is performed. The DNN model is used and further optimized using Talos.

III. SELECTION OF METHOD

While selecting the method of implementation, it is necessary to analyze the advantages and disadvantages of the algorithm according to the use case.

Logistic regression determines the probability of an event occurring given certain predictor variables. This makes it a suitable algorithm for determining CVD risk. It is also simple and computationally efficient, making it easier to apply and interpret. It also allows for the identification of the strength and direction of the association between the predictor variables and the outcome variable. This can help clinicians and researchers better understand the underlying mechanisms of CVD. It is well suited to estimate risk scores that can be used for clinical decision-making and intervention planning.

Logistic Regression models the relationship between an outcome variable and independent variables. The independent variables for this use case will be the risk factors, like age, gender, cholesterol levels, blood pressure, resting heart rate, blood sugar levels, etc. Age, blood pressure, cholesterol levels come under continuous predictor variables, as they can take any value within a particular range. Gender, history of disease falls under categorical predictor variables as they only take on a limited number of values. For this use case, since both categorical and continuous predictor variables are used, logistic regression is well-suited.

The main challenge when it comes to CVD risk assessment is the complexity of algorithms which makes it difficult for healthcare practitioners to interpret the results. Some systems require the personnel to go through additional training to use the risk assessment system, which can be time consuming. Logistic Regression is straightforward and simple, making it easy to interpret. Hence, it is well-suited for clinical settings. It also has an established framework, which makes it very practical and reliable.

IV. DATASET

The parameters used in this study for CVD risk assessment include age, sex, resting blood pressure, cholesterol range, fasting blood sugar, rest electrocardiogram(ECG), maximum heart rate achieved, exercise induced angina, ST depression due to exercise relative to rest, peak exercise ST, range of vessels shown in x-ray, constrictive pericarditis, and thalassemia. Rest ECG is used to measure the electrical activity of the heart. It is crucial to diagnose heart diseases like coronary artery disease and heart attacks

Age is an important predictor variable since it is a well-known risk factor for CVD. The risk increases as the age of the patient increases. Sex is also an important predictor variable, as there are differences in prevalence of CVD between males and females. Resting blood pressure is a predictor of hypertension, which is a major risk factor. Cholesterol range and fasting blood sugar are important indicators of lipid and glucose metabolism respectively. Both of these are known to contribute to the development of CVD.

Maximum heart rate achieved is an important indicator of exercise capacity, which is a strong predictor of CVD outcomes. ST depression due to exercise relative to rest and peak exercise ST are measures of heart function during exercise. Abnormalities in these measures are indicators of underlying CVD. The range of vessels shown in X-ray is an indicator of the extent of coronary artery disease. This is also a key contributor to CVD. Constrictive pericarditis is an inflammation of the pericardium. Pericardium is the covering of the heart. Thalassemia is a genetic disorder that can lead to anemia. This is also a CVD risk factor.

In summary, the parameters that have been used are all important factors that contribute to the development and progression of CVD. Inclusion of these variables in the predictive model makes it possible to identify individuals at an increased risk for CVD, making it easier to provide them with preventative measures and treatment.

V. METHODOLOGY

Cardiovascular Disease Detection Systems are used in a lot of institutions such as hospitals, research institutions and public health agencies. Some of them still rely on using manual decision making or rule-based systems. Some also rely on statistical methods or basic machine learning algorithms such as naïve Bayes classifiers or decision trees. Logistic Regression can handle a vast range of continuous and categorical predictor variables. It is also able to model relationships that are non-linear, as the data in healthcare can be complex. The results provided by logistic regression are also easy to interpret, which is important as it facilitates easier understanding of the factors contributing to the disease.

This systemuses logistic regression for the prediction of cardiovascular disease. A CSV file containing the 13 attributes ranging from gender, age, fasting blood sugar, rest ECG etc., are used as the dataset to train the model. Using these attributes, the model predicts the probability of the response variable. The aim of this system is to improve accessibility, so the view layer is incorporated using Django. Django is a python framework which is powerful and flexible. This has been used to incorporate the machine learning model as a web application.

VI. ARCHITECTURE OF THE SYSTEM

A multi-tiered approach has been followed for the architecture of the system. This approach consists of three tiers - the data tier, presentation tier and logic tier.

For the data tier, medical records of patients containing 13 attributes were used. These attributes were age, sex, resting blood pressure, cholesterol range, fasting blood sugar, rest ECG, maximum heart rate achieved, exercise induced angina, ST depression due to exercise relative to rest, peak exercise ST, range of vessels shown in x-ray, constrictive pericarditis and thalassemia. Constrictive pericarditis is the thickening of the covering of the heart. Thalassemia is a disorder where the quantity of hemoglobin produced is insufficient. Some attributes are continuous predictor variables and some are categorical predictor variables. This labeled dataset is stored in the form of a csv file.

Logistic Regression algorithm is used to predict CVD risk using the data that the user inputs. This is the business logic of the application. The model is trained using the dataset consisting of attributes from various patients. It learns by finding the coefficients that relate the independent variables (input) to the dependent variables (output). The model operates using a sigmoid function with an output of a probability score between 0 to 1.

The trained logistic regression model is deployed using the web framework Django. The Presentation tier of the system comprises the user interface. The web application is a form that takes input from the user. There are 13 input variables and each input variable accepts a value within the range for the said attribute. The web application is also responsible for displaying the prediction results. The presentation tier is implemented using frontend technologies like HTML, CSS and JavaScript.

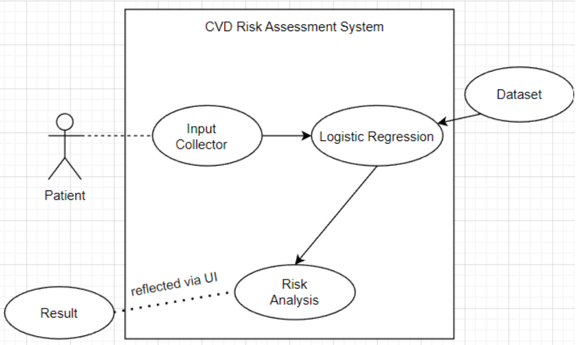

Fig. 1. Basic architecture of system

VI. IMPLEMENTATION

A dataset containing attributes that can be used for the risk assessment of CVD is collected and preprocessed. The dataset is then split into two – a training set and a testing set. The model's performance can be analyzed this way. The logistic regression model is trained using the training set. It is implemented using the Scikit-learn library. The model is then deployed using Django to create a web application. The user interface of the web application is implemented using HTML, CSS and JavaScript.

The home page consists of a quick predict feature where users can input the required fields and the web application returns the probability of CVD. It also has an option to login or signup. Users can sign up by giving their basic details such as username, email and password. These details can be used to login to the system. Users who are logged in can access the user dashboard which has several features. The user's previous records are saved if they used the web application while logged in. Hence, the user can compare all their previous records to see if the risk of cardiovascular disease has reduced or increased.

Once the user has completed checking their risk, there is also an option to see the list of all hospitals and doctors available so they can contact them if they are in need of further medical assistance. The other features in the dashboard include symptoms and preventative measures. The symptoms feature helps people understand the risk of cardiovascular diseases and know when urgent medical care is needed. The preventative measures feature has a list of general tips and practices that can be incorporated into one's lifestyle to reduce the risk of cardiovascular disease. There also is an admin dashboard that administrators can access. The admin dashboard consists of a list of all users in the system and their details. The administrator has permissions to add or delete users. The administrator also can add the list of doctors and their relevant details that can be seen by a user.

VII. COMPARATIVE ANALYSIS

Several algorithms such as K-Nearest Neighbors, Logistic Regression, Naive Bayes, Support Vector Machine were compared to determine their accuracies.

K-Nearest Neighbors is a machine learning algorithm that mainly uses the similarity of a feature for classifying the data. According to the similarity of a point in the dataset, the new data will be assigned a value. The accuracy scores are plotted against the different values of K using the Matplotlib library. Using the dataset, the accuracy was measured to be 78.69%.

Logistic Regression model is a supervised machine learning model. The output is a range between 0-1 which represents the probability of the outcome variable being true. The accuracy score was generated using the score() function from scikit-learn library. The score() function takes in the test input data 'x_test' and test output data 'y_test' as inputs and returns the mean accuracy on the given test data and labels. Using the dataset, the accuracy was measured to be 86.89%.

Naive Bayes is a classification model. It uses the assumption that all the predictors are unrelated to each other. It makes its predictions by using the Bayes theorem. The Bayes theorem states that the probability of one event that's related to another event occurring, is the probability of the second event given the occurrence of the first, multiplied by the probability of the first event. The score() method is called to obtain the accuracy score of the model on the test data. The values that are predicted are compared to the actual values in the test set and the percentage of correctly classified instances is calculated. Using the dataset, the accuracy was measured to be 85.25%.

Support Vector Machine is a supervised learning algorithm that is used for both classification and regression. SVM divides the data into two sections separated by a hyperplane. The hyperplane is a space that divides the data into two separate classes. The score() method was used to evaluate the performance of SVM on the test data (x_test and y_test). Using the dataset, the accuracy was measured to be 81.97%.

All the accuracy scores were added to a dictionary called 'accuracies'. The data in this dictionary is used to generate a bar plot that compares the accuracy of different machine learning algorithms. Each algorithm is represented by a bar and the height of the bar represents the accuracy achieved by the corresponding algorithm. The x-axis indicates the different algorithms used, whereas the y-axis indicates the accuracy in percentage. The plot was created using the seaborn library, and the style was set to whitegrid. The x-label was set to 'Algorithm' while the y-label was set to 'Accuracy (%)'.

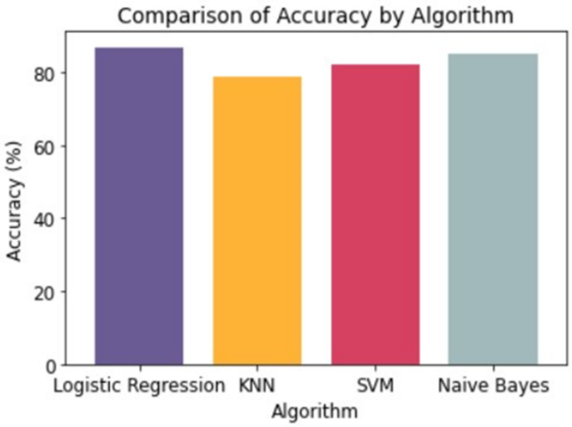

Fig. 2. Graph Comparing Accuracies of Algorithms

VIII. RESULT

The advantage of implementing the system with Logistic Regression can be analyzed using accuracy as the metric. Algorithms such as K-Nearest Neighbors, Logistic Regression, Naive Bayes, Support Vector Machine were tested to determine the accuracies.

After comparison, it can be deduced that the system using Logistic Regression for assessing CVD risk is more accurate than the other algorithms. The comparison is done by testing the accuracy of different models using a separate dataset which is labeled.

Logistic Regression is also advantageous as it is straightforward and simple.It is also computationally efficient. It is easy to interpret - making it suitable for clinical settings. It also has a well-established framework, which makes it very practical and reliable. The logistic regression model is also more efficient than manual systems, as it reduces the possibility of human errors.

IX. CONCLUSION

The accuracy of using Logistic Regression to assess CVD risk is higher than the other systems. As the predictions are consistent and objective, it reduces the chances of human error due to subjectivity. It can also handle large amounts of data and is scalable while also being easy to interpret, making it fit for the healthcare system. Cardiovascular disease is one of the highest causes of death in the world. The main objective of this study was to identify individuals at high risk and initiate further treatment if needed, thus helping the current healthcare system become more efficient.

X. REFERENCES

[1] Jaymin Patel, Tejal Upadhyay, Dr. Samir Patell, Heart Disease Prediction Using Machine Learning and Data Mining Technique, IJCSC, Vol 7 (2016).

[2] V.V. Ramalingam, Ayantan Dandapath, M Karthik Raja, Heart-disease prediction using machine learning techniques, International Journal of Engineering and Technology (2018).

[3] Vinitha S, Sweetlin S, Vinusha H and Sajini S, Disease Prediction Using Machine Learning Over Big Data, Computer Science & Engineering: An International Journal (CSEIJ) (2018).

[4] Pooja Anbuselvan, Heart Disease Prediction using Machine Learning Techniques, International Journal of Engineering Research & Technology (2020).

[5] Md. Ehtisham Farooqui, Dr. Jameel Ahmad.: Disease Prediction System using Support Vector Machine and Multilinear Regression, International Journal of Innovative Research in Computer Science & Technology (2020).

[6] Palle Pramod Reddy, Dirisinala Madhu Babu, Hardeep Kumar, Dr.Shivi Sharma, Disease Prediction using Machine Learning, International Journal of Creative Research Thoughts (2021).

[7] Raj H. Chauhan, Daksh N. Naik, Rinal A. Halpati, Sagarkumar J. Patel, Mr. A.D.Prajapati, Disease Prediction using Machine Learning, International Research Journal of Engineering and Technology (2020).

[8] Tarigoppula V.S Sriram, M. Venkateswara Rao, G V Satya Narayana, DSVGK Kaladhar, T Pandu Ranga Vital, Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms, International Journal of Engineering and Innovative Technology (2013).

[9] Shahadat Uddin, Arif Khan, Md Ekramul Hossain, Mohammad Ali Moni, Comparing different supervised machine learning algorithms for disease prediction, BMC Medical Informatics and Decision Making (2019).

[10] Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivatsava, Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques, Institute of Electrical and Electronics Engineers (2019).

[11] Pankaj Mathur, Shweta Srivastava, Xiaowei Xu, Jawahar L Mehta, Artificial Intelligence, Machine Learning, and Cardiovascular Disease, Clinical Medicine Insights, Cardiology (2020).

[12] Yankun Cao, Zhi Liu, Pengfei Zhang, Yushuo Zheng, Yongsheng Song, Lizhen Cui, Deep Learning Methods for Cardiovascular Diseases, Journal of Artificial Intelligence and System (2019).

[13] A.V. SITAR-TĂUT, D. ZDRENGHEA, D. POP, D.A. SITAR-TĂUT, Using Machine Learning Algorithms in Cardiovascular Disease Risk Evaluation, Journal of Applied Computer Science & Mathematics (2009).

[14] An Dinh, Stacey Miertschi, Amber Young, and Somya D. Mohanty, A data-driven approach to predicting diabetes and cardiovascular disease with machine learning, BMC Medical Informatics and Decision Making (2019).

[15] Sumit Sharma, Mahesh Parmar, Heart Diseases Prediction using Deep Learning Neural Network Model, International Journal of Innovative Technology and Exploring Engineering (2020).