Cardiovascular Prediction System Using Machine Learning

Pooja Gole[1], Akshada Dhumal[2], Piyush Jadhav[3]

Department of Computer Engineering, Narhe

Abstract

A heart attack is another term for a cardiovascular stroke. According to the World Health Organization, stroke is the second leading cause of death worldwide, accounting for 11% of all deaths. As a result of the high number of COVID-19 patients who experience breathing issues as a side effect of therapy, the risk of a stroke has increased. There isn't always a system in place to check for the odds of having a stroke. A modest fluctuation in blood pressure that we underestimate can sometimes lead to a stroke. So, we created a system that will predict the possibility of having a stroke. We experimented with various of machine learning techniques to come up with the optimal of the solution. We chose the best performing algorithm in terms of accuracy and type-1 errors among the majority of machine learning algorithms that we tested with implementation. We devised a method for predicting the likelihood of suffering a stroke based on a few simple characteristics that maybe measured at home. As a result, if our algorithm predicts that you may have a stroke, you should seek medical advice as soon as possible. The chosen model is trained using a training set and evaluated using appropriate metrics such as accuracy, precision, recall, and AUC-ROC. The final model is deployed for predicting heart attacks on new, unseen data. The results demonstrate promising performance, with high accuracy and reliable prediction capabilities. The developed model holds significant potential for assisting healthcare professionals in early detection and prevention of heart disease, thereby improving patient care and reducing mortality rates. Future work may focus on expanding the dataset, incorporating additional features, and exploring advanced machine learning techniques to further enhance the predictive capabilities of the model.

Keywords: AWS, Cardiovascular Disease Prediction, Machine Learning Techniques, Random Forest model.

I. INTRODUCTION

According to the World Health Organization, every year 12 million deaths occur worldwide due to heart disease. Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of data analysis. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduces the complications. Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the health care industry. This project aims to predict future heart disease by analysing data of patients which classifies whether they have heart disease or not using machine-learning algorithm. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analysing to extract the desired data we can say that this technique can be very well adapted to do the prediction of heart disease.

II. LITERATURE SURVEY

1.The paper Cardiovascular Disease Prediction Using Machine Learning Models was published in dec 2020 by IEEE. In this system decision tree classifier is used. the decision tree classifier shows quite good performance as compared to all other models where the performance of a model is evaluated in terms of classification accuracy. Various studies are present which focuses on heart disease prediction where the diagnosis is done by using different techniques of data mining. [1]As per the research done by the research group in the decision tree classifier shows quite good performance as compared to all other models where the performance of a model is evaluated in terms of classification accuracy. [2] In the focus is on developing a system to help medical professionals to evaluate the risk of heart disease of a patient based on the patient’s clinical data. [3] In Heart disease prediction is done using machine learning where the parameters used are

Age, Sex, Blood Pressure, Heart Rate, Diabetes, Hyper cholesterol, Body Mass Index (obesity). In ANN, KNN, k-means, and K-medoids algorithms are trained on the level and dataset for heart disease.

2.The paper Cardiovascular Disease Forecast using Machine Learning Paradigms was published in April 2020. In this system, Logistic regression, Naive Bayes, SVM, and Decision Tree classification algorithm are used. [1] This section is for presenting the research demand on this topic and some works that must highlight our study. We found that a lot of research was focused on cardiovascular disease. We were keen to know the risk factors for predicting heart disease. [6] Nabaouia Louridi, Meryem Amar and Bouabid El Ouahidi used different machine learning algorithms to identify the cardiovascular disease where they finally proposed a SVM with a linear kernel approach. In this approach they used features and found an accuracy of 66.8%. In 2019 N. Satish Chandra Reddy, Song Shue Nee, Lim ZhiMin & Chew Xin Ying stated that Random Forest can be used as classification algorithm to train the system for identifying CVD with 70% to 75% accuracy.[3] They used 14 features Using the Dataset of UCI library in 2018 Aditi Gavhane, Gouthami Kokkula, Isha Pandya & Prof. Kailas Devadkar conducted a study to predict the heart disease of a human using 13 features among 76. In this study they used MLP and got an average of 0.91 precision.

3. The paper Cardiovascular Disease Prediction using Machine Learning Algorithms was published in June 2020. In this system uses Convolutional neural network for predicting the disease risk. They also followed the generic way of doing data analysis.

4. The paper A Method of Cardiovascular Disease Prediction using Machine Learning was published in July 2021. The heart disease prediction can be carried out using various algorithms such as Support Vector Machine classifier, decision tree.

5. The paper Heart disease prediction using machine learning technique was published in April 2020. In this system, the input details are obtained from the patient.[1] Then from the user inputs, using ML techniques heart disease is analysed. Now, the obtained results are compared with the results of existing models within the same domain and found to be improved.[2] The data of heart disease patients collected from the UCI laboratory is used to discover patterns with NN, DT, Support Vector machines SVM, and Naive Bayes. The results are compared for performance and accuracy with these algorithms.[3] The proposed hybrid method returns results of 67% for F-measure, competing with the other existing methods.

6. The paper Heart disease prediction using machine algorithm was published in October 2020 Day by day the cases of heart diseases are increasing at a rapid rate and it's very Important and concerning to predict any such diseases beforehand. [1] This diagnosis is a difficult task i.e. it should be performed precisely and efficiently. The research paper mainly focuses on which patient is more likely to have a heart disease based on various medical attributes. [7] We prepared a heart disease prediction system to predict whether the patient is likely to be diagnosed with a heart disease or not using the medical history of the patient. We used different algorithms of machine learning such as logistic regression and KNN to predict and classify the patient with heart disease. [3] A quite Helpful approach was used to regulate how the model can be used to improve the accuracy of prediction of Heart Attack in any individual.[4] The strength of the proposed model was quiet satisfying and was able to predict evidence of having a heart disease in a particular individual by using KNN and Logistic Regression which showed a good accuracy in comparison to the previously used classifier such as naive bayes etc. [5] So a quiet significant amount of pressure has been lift off by using the given model in finding the probability of the classifier to correctly and accurately identify the heart disease. The Given heart disease prediction system enhances medical care and reduces the cost. [6] This project gives us significant knowledge that can help us predict the patients with heart disease It is implemented on the .pynb format.

III. OBJECTIVES OF THE PROJECT

1. Predicted the possibility of getting a stroke.
2. Learn the implementation of different machine learning algorithms.

IV. PROBLEM STATEMENT

The Prediction of Cardiovascular Stroke is a classification problem. We must classify the provided data across multiple classes based on patterns in the input attributes in a classification issue. "Stroke" and "no stroke" are the two classifications available in our problem. We used multiple classification machine learning algorithms, as well as numerous sets of trials based on various feature extraction methodologies and algorithms to category.

V. METHODOLOGY

1.Data Preprocessing

A component of data preparation, describes any type of processing performed on raw data to prepare it for another data processing procedure. It has traditionally been an important preliminary step for the data mining process. More recently, data Preprocessing techniques have been adapted for training machine learning models and AI models and for running inferences against them. Data Preprocessing transforms the data into a format that is more easily and effectively processed in data mining, machine learning and other data science tasks. The techniques are generally used at the earliest stages of the machine learning and AI development pipeline to ensure accurate results. There are several different tools and methods used for Preprocessing data, including the following:

- sampling, which selects a representative subset from a large population of data;
- transformation, which manipulates raw data to produce a single input;
- denoising, which removes noise from data;
- imputation, which synthesizes statistically relevant data for missing values;
- normalization, which organizes data for more efficient access; and
- feature extraction, which pulls out a relevant feature subset that is significant in a particular context.

Exploratory Data Analysis (EDA)

It is a way of visualizing, summarizing, and interpreting the information that is hidden in rows and column format. EDA is one of the crucial steps in data science that allows us to achieve certain insights and statistical measure that is essential for the business continuity, stakeholders, and data scientists. It performs to define and refine our important features.

Understanding Data
Handle Missing value
Removing duplicates
Outlier Treatment
Normalizing and Scaling
Encoding Categorical variables

Bivariate Analysis

. 2.Model building

The modelling process was divided into two main parts: Building different model of machine learning to find out best algorithm and hyperparameter optimization of selected model.

Random Forest Classifier

Random forests or random decision forests is a learning method for classification and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes become our model's prediction. The fundamental concept behind random forest is a simple but powerful one the wisdom of crowds.

In data science speak, the reason that the random forest model work. Many relatively uncorrelated models operating as a committee will outperform any of the individual constituent models. Random forests are frequently used as "Blackbox" models in businesses, as they generate reasonable predictions across a wide range of data while requiring little configuration. Before understanding the working of the random forest, we must investigate the ensemble technique. Ensemble simply means combining multiple models. Thus, a collection of models used to make predictions rather than an individual model. Ensemble uses two types of methods i.e. Bagging creates a different training subset from sample training data with replacement the final output is based on majority voting. For example, Random Forest. Boosting combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control overfitting. The sub-sample size is controlled with the max samples parameter if bootstrap=True (default).

. 3.Evaluation metrics

Accuracy

Accuracy is one metric for evaluating classification models. Informally, Accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

(1)

Precision And Recall

Precision is a good measure to determine, when the costs of False Positive is high. For instance, email spam detection. In email spam detection, a false positive means that an email that is non spam (actual negative) has been identified as spam (predicted spam). The Email user might lose important emails if the precision is not high for the spam detection model. In the field of information retrieval precision is the fraction of retrieved documents that are relevant to the query.

$$\text{Precision} = \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Positive}}$$

$$= \frac{\textit{True Positive}}{\textit{Total Predicted Positive}}$$

(2)

Recall

Recall calculates how many of the Actual Positives our model capture through labelling it as True Positive. Applying the same understanding, we know that

Recall shall be the model metric we use to select our best model when there is a high cost associated with False Negative.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$= \frac{True\ Positive}{Total\ Actual\ Positive}$$

(3)

F1 Score

F1 is a function of Precision and Recall. F1 Score might be a better measure to use if we need to seek a balance between Precision and Recall and there is an

uneven class distribution (large number of Actual Negatives).

$$\text{F1} = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

(4)

Support

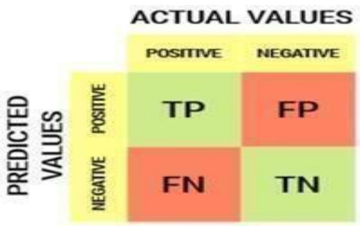
Support is the number of actual occurrences of the class in the specified data set. Imbalanced support in the training data may indicate structural weaknesses

in the reported scores of the classifier and could indicate the need for stratified sampling or re-balancing.

Confusion Matrix

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making.

□ The target variable has two values: Positive or Negative □ The columns represent the actual values of the target variable □ The rows represent the predicted values of the target variables.



		ACTUAL VALUES	
		POSITIVE	NEGATIVE
PREDICTED VALUES	POSITIVE	TP	FP
	NEGATIVE	FN	TN

Fig.Confusion Matrix

4.PYTHON DJANGO

Django is a high-level Python web framework that encourages rapid development and clean, pragmatic design. Built by experienced developers, it takes care of much of the hassle of web development, so you can focus on writing your app without needing to reinvent the wheel. It is free and open source. A Web framework is a set of components that provide a standard way to develop websites fast easily. Django primary goal is to ease the creation of complex database driven websites. Some well-known sites that use Django include instagram, disqus, washington times, bitbucket and mozilla. Django follows the Model View Template Pattern design. The model provides data from the database. In Django, the data is delivered as an Object Relational Mapping, which is a technique designed to make it easier to work with databases. The most common way to extract data from a database is structured query language. One problem with structured query language is that you must have a pretty good understanding of the database structure to be able to work with it. Django, with object relational database, makes it easier to communicate with the database, without having to write complex structured query language statements. The models are usually located in a file called models.py. A view is a function or method that takes http requests as arguments, imports the relevant model, and finds out what data to send to the template, and returns the result. The views are usually located in a file called views.py. A template is a file where you describe how the result should be represented. Templates are often .html files, with HTML code describing the layout of a web page, but it can also be in other file formats to present other results, but we will concentrate on .html files.

Design Philosophies of Django

- Loosely Coupled: Django aims to make each element of its stack independent of the others.
- Less Coding: Less code so in turn a quick development.
- Do not Repeat Yourself: Everything should be developed only in exactly one place instead of repeating it again and again.

VI. CONCLUSION

In conclusion, the heart attack prediction project aimed to develop a model that could accurately predict the occurrence of heart attacks based on relevant clinical parameters and features. The project involved steps such as data collection, Preprocessing, exploratory data analysis, model selection, training, evaluation, and deployment. Through thorough analysis and modelling, a suitable machine learning algorithm, such as the Random Forest Classifier, was selected and trained on the dataset. The model demonstrated promising performance in predicting heart attacks.

VII. ACKNOWLEDGEMENT

First and foremost, we extend our sincere appreciation to the healthcare professionals and researchers who have dedicated their time and expertise in collecting and providing the valuable dataset used in this project. Their efforts have been instrumental in enabling us to conduct meaningful analysis and develop a predictive model.

VIII. REFERENCE

[1] "Cardiovascular diseases" Available: https://www.who.int/en/newsroom/fact-sheets/detail/cardiovascular- diseases. [Accessed: 25- January- 2020]

[2] Wu, Ching-seh Mike, Mustafa Badshah, and Vishwa Bhagwat, "Heart Disease Prediction Using Data Mining Techniques." In Proceedings of the 2019 2nd International Conference on Data Science and Information Technology, pp. 7-11. 2019.

[3] "What should my heart rate be?" https://www.medicalnewstoday.com/articles/235710.php#normalresting- heart-rate. [Accessed: 21- January- 2020]

[4] Nabaouia Louridi, Meryem Amar, Bouabid El Ouahidi "IDENTIFICATION OF CARDIOVASCULAR DISEASES USING MACHINE LEARNING", 7th Mediterranean Congress of Telecommunications (CMT), 2019.

[5] N. Satish Chandra Reddy, Song Shue Nee, Lim Zhi Min & Chew Xin Ying "Classification and Feature Selection Approaches by Machine Learning Techniques: Heart Disease Prediction", International Journal of Innovative Computing, 2019,

[6] Aditi Gavhane, Gouthami Kokkula, Isha Pandya & Prof. Kailas Devadkar (PhD) "Prediction of Heart DiseaseUsing Machine Learning", ICECA 2018, IEEE Xplore.

[7] Sonakshi Harjai1 & Sunil Kumar Khatri, "An Intelligent Clinical Decision Support System Based on ArtificialNeural Network for Early Diagnosis of Cardiovascular Diseases in Rural Areas", AICAI, 2019, DOI: 10.1109/AICAI.2019.8701237.

[8] Senthil Kumar Mohan, Chandrasegar Thirumalai and Gautam Srivastava, "Effective Heart Disease PredictionUsing Hybrid Machine Learning Techniques", Special Section on Smart Caching, Communications, Computing and Cybersecurity For Information-centric Internet Of Things, IEEE Access Volume 7, 2019.

[9] Duraipandian, M. "Performance Evaluation of Routing Algorithm for MANET based on the Machine Learning Techniques." Journal of trends in Computer Science and Smart technology (TCSST) 1, no. 01 (2019): 25-38.

[10] "Dataset" Available: https://github.com/istyak/hd/blob/master/a4.csv. [Accessed: 12- Dec- 2019].

[11] Suthaharan S., "Support Vector Machine. In: Machine Learning Models and Algorithms for Big Data Classification". Integrated Series in Information Systems, vol 36. Springer, Boston, MA, 2016.

[12] Arundhati Navada, Aamir Nizam Ansari, Siddharth Patil, Balwant A. Sonkamble, "Overview of Useof Decision Tree algorithms in Machine Learning". IIEEE Control and System Graduate Research Colloquium