INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 9, Issue 5 - V9I5-1186)
Available online at: https://www.ijariit.com

Care: Cardiac attack risk estimation using Machine Learning

Patan Imran Khan
p.imrankhan3579@gmail.com
Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh

Kothamasu Surya Ratna
20BQ1A1287@vvit.net
Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh

Meda Gopi Krishna
20BQ1A12A8@vvit.net
Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh

Nutalpati Ashok
nutalpati.ashok@vvit.net
Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh

ABSTRACT

We are revolutionizing heart attack risk assessment with our ground-breaking initiative, "CARE: Cardiac Attack Risk Estimation Using Machine Learning," by utilizing machine learning models' predictive power. Our technology uses past data analysis to forecast the likelihood of a subsequent heart attack based on user-supplied details such as physical attributes, symptoms, and medical background. Our project's main goal is to reduce the burden on the healthcare system by providing users with remote access to screening facilities that can identify people at both low and high risk. With the goal of improving the precision of heart attack risk predictions, our ground-breaking platform, the "Heart Attack Risk Predictor," is a groundbreaking venture into the field of machine learning.

Keywords: *Revolutionizing, Forecast, Data analysis, Precision, Screening Facilities.*

1.INTRODUCTION

Heart attacks continue to be the most common and deadly cause of death worldwide, taking more lives than any other illness. According to data released by the World Health Organization (WHO), around 18 million people died from heart disease in 2016, accounting for a startling 30% of all deaths globally. Developing and impoverished countries were mostly responsible for this load.
When blood supply to the heart is entirely restricted or severely decreased, a heart attack happens. The accumulation of fat and cholesterol in the coronary arteries, which results in the creation of plaques and is known medically as atherosclerosis, is the cause of this frequently deadly blockage.

Heart attacks, also known as coronary heart disorders, are dangerous and common. There is a heart attack death in the United States alone.Making healthy dietary and exercise choices, as well as giving up smoking, can dramatically lower the chance of having a heart attack, especially when paired with prompt medical intervention. However, a number of contributing factors, including high blood pressure, diabetes, and high cholesterol, make it difficult to identify high-risk people.Due to its capacity to recognize patterns and classifications, machine learning algorithms are being employed more and more in the medical industry to create screening tools.
Throughout the world, cardiovascular illnesses remain a primary cause of morbidity.

In the medical field, risk prediction is a challenging undertaking that requires a great deal of data and medical knowledge. Automated medical diagnostic systems can lower expenses and increase efficiency.
Quickly determining a patient's risk level based on health-related

2.LITERATURE SURVEY

1. KNN for Predicting Heart Disease by Agung Enriko et al.
Heart disease prediction parameters were reduced by Enriko and colleagues to just 8, making them appropriate for M2M remote patient monitoring.They obtained accurate results by using KNN with parameter weighting.Using 8 parameters with KNN outperformed using 13 parameters with KNN, Naive Bayes, and Decision Tree in terms of accuracy.

2. Genetic algorithms and neural networks by Ashwini Shetty and Chandra Naik
A technique for predicting cardiac disease was created by Shetty and Naik.They employed genetic algorithms and neural networks, preprocessed the data, and used 13 risk factors from a dataset.
Taking into account all 13 characteristics, the system used a history cardiac database to diagnose the patient.

3. Analysis of Cardiovascular Disease by Sultana, Haider, and Uddin
The study concentrated on data mining techniques that can help medical practitioners predict heart disease.
By timing how long it took to produce a decision tree, they evaluated the performance of the model.The idea was to use fewer characteristics to predict disease.

4.Kirmani: Data Mining Methods for Prognosticating Illnesses
Kirmani used data mining to predict a range of illnesses and cut down on physical examinations.The Cleveland database's features were increased to fifteen by applying decision trees and neural networks.

5. Cleveland Heart Database Expansion: Chaitrali and Sulabha
These researchers added obesity and smoking as two more characteristics to the frequently used Cleveland Heart disease database.In data mining, they employed Decision Trees, Naive Bayes, and Neural Networks for classification.

3.METHODOLOGY

The dataset originates from the Massachusetts city of Framingham, where There was research being done on the heart. The Kaggle website makes it accessible to the general public. This project's goal is to determine the risk involved in developing future years will see an increase in coronary heart disease (CHD). if there is a chance the patient could develop heart disease (CHD) during the next ten years. The information gathered includes details regarding the patients. There are fifteen qualities total and nearly four thousand records. Every feature has the potential to turn into a risk factor

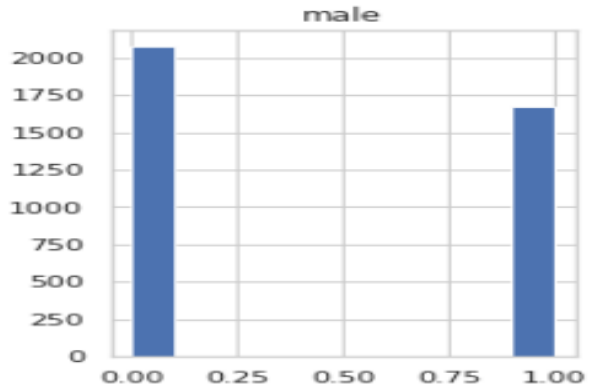
3.1 Attributes

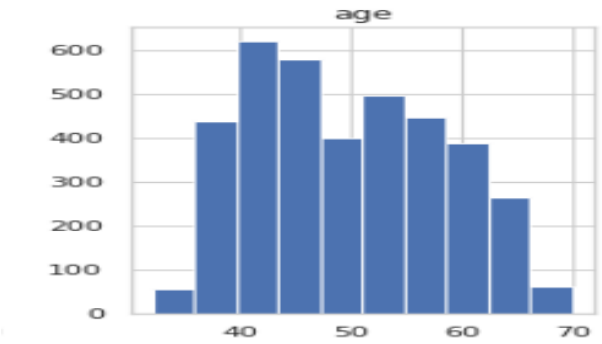
The attributes in a dataset related to demographics, behavior, medical history, and current health are described in this section. The qualities are summarized as follows:
Demographic Characteristics:
Sex: The patient's gender, where 0 denotes a female and 1 a male.
Age: Although recorded as an integer, the patient's age is regarded as a continuous variable.
Behavioral Characteristics:
Present Smoker: This binary characteristic indicates if the patient smokes at the moment (0 for no, 1 for yes).
Cigars Per Day: Because fractional values may occur, this measure, which is an average of the number of cigarettes smoked each day, is considered continuous.
Details regarding medical history:
BP Meds: This binary attribute (0 for no, 1 for yes) indicates if the patient was taking blood pressure medication.
Prevalent Stroke: A binary variable that indicates whether or not the patient has had a stroke in the past (0 for no, 1 for yes).
Prevalent Hyp: A binary characteristic with 0 meaning the patient was not hypertensive and 1 meaning that they were.
Diabetes: A binary characteristic (0 for no, 1 for yes) that indicates whether or not the condition exists.
Present Health Status:
Sys BP: The patient's systolic blood pressure.
Dia BP: The patient's diastolic blood pressure.
Tot Chol: The patient's total cholesterol.
Heart Rate: Although heart rate is discrete, it is handled as a continuous variable.
Glucose: The continuous blood glucose level.
Body Mass Index, or BMI, is a continuous variable that shows the relationship between a person's height and weight.
These characteristics offer useful data for evaluating a patient's health and risk factors for different illnesses, such as heart disease. In medical research and healthcare settings, they serve as the foundation for data analysis and predictive modeling.

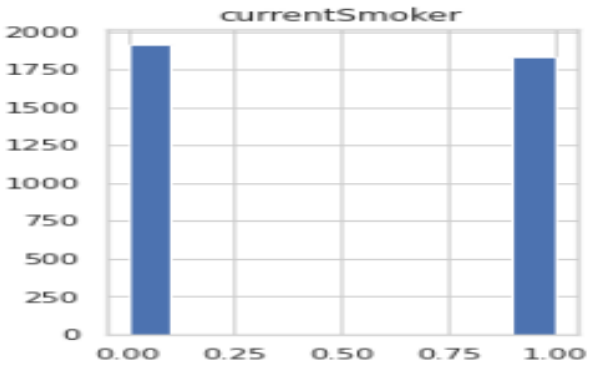
3.2 Data Analysis

In order to obtain important statistical insights from the data, we examined the important odds and proportions for the categorical attributes, as well as the distributions of the various qualities, their interactions with one another, and the objective variable.

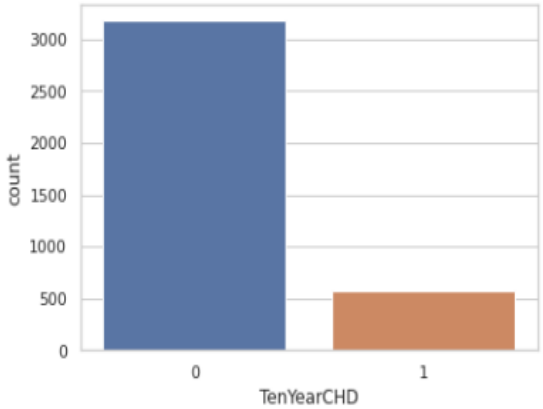
The first step was to look at how different qualities were distributed and determine which was best shown with the aid of histograms.







Categorical and continuous variables are distinguished by the distribution graphs. Not only that, but only a small percentage of patients had diabetes, hypertension, or were taking medication to control their blood pressure. None of the participants had ever experienced a stroke. We examined the number of positive and negative examples to confirm our suspicions that the dataset might be uneven, as indicated by the following plot. There were 572 individuals with CHD and 3179 individuals who were suffering from the condition.



3.4 Handling Imbalance Data

In machine learning and data science, the term "imbalanced data distribution" is often used to characterize situations where observations in one class are significantly higher or lower than observations in other classes. Since the aim of machine learning algorithms is to reduce error and increase accuracy, they do not take into account class distribution. This problem is exemplified by facial recognition, anomaly detection, and fraud detection.

Decision Tree and Logistic Regression, for example, are common machine learning algorithms that favor the majority and ignore the minority. They tend to predict just the majority class, which leads to the false predictions of the minority class. Technically speaking, if our dataset has an uneven data distribution, our model is more susceptible to the circumstance where the minority class has a negligible or extremely low recall. Unbalanced Data

Handling Techniques: There are two main techniques for dealing with an unequal distribution of classes.

1. SMOTE
2. Near Miss Algorithm

4.Experimental Results

We have used the confusion matrix for evaluating the accuracy, specificity, and sensitivity of our model. We have chosen this to get more idea about the prediction nature of our model. In other words, we can compare the sensitivity and specificity to have an idea of whether our model is predicting true positives more accurately or it is predicting the true negative more accurately. In the end, we can see the overall accuracy of our model.

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

4.1 Confusion Matrix

Predicted 1 = positive,
Predicted 0 = negative,
TN = true negative.,
TP = true positive,
FP = false positive,
FN = false negative.

5. CONCLUSION

More precise and accurate disease prediction methods are desperately needed in the modern world, given the startling rise in deaths linked to cardiovascular diseases, particularly heart attacks. The primary aim of this project is to create a method that can accurately predict an individual's risk of heart disease over a period of ten years. Numerous methods of data analysis and machine learning have been used to accomplish this.

The k-nearest neighbor (k-NN) algorithm, also known as the "lazy" algorithm because of its simple and data-driven methodology, and the traditional nearest neighbor have both been thoughtfully included into the project. These techniques are essential for finding patterns and connections in the data that can help make disease predictions more precise.

Boruta, a sophisticated variant of the random forest technique, has been used to carry out feature selection, a crucial step in improving model performance and interpretability. The most informative features from the dataset should be carefully chosen because they have a big influence on prediction accuracy.

The UCI repository is a reliable source of machine learning datasets, and the project has used its resources to obtain the dataset. The reliability and consistency of the data used for analysis are guaranteed when pre-existing datasets are leveraged.

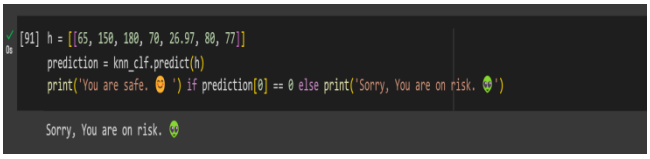
Let’s use a random person as an example for the test. A 39-year-old man with 195 total cholesterol, 106 systolic BP, 70 diastolic BP, a BMI of 26.97, a heart rate of 80, and glucose levels of 77. Data sent to the backend for this particular person will be

[39, 195, 106, 70, 26.97, 70, 77]

```
# age totChol sysBP diaBP BMI heartRate glucose
h = [[39, 195, 106 ,70 , 26.97, 70, 77]]
prediction = knn_clf.predict(h)
print('You are safe. ') if prediction[0] == 0 else print('Sorry, You are on risk. ')
```

```
You are safe.
```


Let's use a random person as an example for the test. A 65-year-old man with 150 total cholesterol, 180 systolic BP, 70 diastolic BP, a BMI of 26.97, a heart rate of 80, and glucose levels of 77. Data sent to the backend for this particular person will be
[65, 150, 180, 70, 26.97, 80, 77]



5.2 FUTURE WORK

This project's scope presents enormous opportunities for future growth and improvement. The development of an intuitive user interface that enables people to engage with the prediction model is a crucial path for growth. Users could enter their personal health information to get an estimate of how likely they were to have a heart attack. This gives the public and medical professionals a useful tool in addition to encouraging a proactive approach to healthcare.

Furthermore, more investigation into the connection between age, gender, and the risk of heart attacks may be undertaken in the future. Comprehending the demographic patterns associated with heart disease can facilitate the customization of preventive and intervention tactics.

The quality of the dataset presented a significant challenge for the project. Improving data reliability in the future will require sourcing from more reliable and extensive databases. Predictive models' practical value can be strengthened by substantially increasing their accuracy through data that comes from a reliable source.

Data preprocessing and normalization techniques can be used to reduce the effects of overfitting and underfitting in order to further improve model performance. To guarantee that the predictive models are robust and that they can effectively generalize to new data, proper data handling is crucial.

As this project develops, more machine learning models will be able to be assessed and different datasets will be combined for comparative analysis. By using this method, one can improve the predictive accuracy and identify the models that work best. The ultimate objective is to detect heart attack risks early on, giving people the chance to take preventative action and enhancing their general health. Future iterations of this project have the potential to significantly advance personalized medicine and cardiovascular health.

In order to give real-time data inputs for the predictive model, the project can also investigate the integration of cutting-edge health monitoring technologies, such as wearables and electronic health records. This stream of dynamic data may provide more current and individualised risk assessments.

6. REFERENCES

[1] Enriko, I. K. A., Suryanegara, M., & Gunawan, D. (2016). Heart disease prediction system using k-Nearest neighbor algorithm with simplified patient's health parameters. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 8(12), 59-65.

[2] Shetty, A., & Naik, C. (2016). Different data mining approaches for predicting heart disease. Int J Innov Res Sci Eng Technol, 5(9), 277-281.

[3] Sultana, M., Haider, A., & Uddin, M. S. (2016, September). Analysis of data mining techniques for heart disease prediction. In 2016 3rd international conference on electrical engineering and information communication technology (ICEEICT) (pp. 1-5). IEEE.

[4] Kirmani, M. M. (2017). Cardiovascular disease prediction using data mining techniques: A review. Oriental Journal of Computer Science & Technology, 10(2), 520-528.

[5] Dangare, C. S., & Apte, S. S. (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-48.

[6] Melillo, P., De Luca, N., Bracale, M., & Pecchia, L. (2013). Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability. IEEE journal of biomedical and health informatics, 17(3), 727-733.

[7] Guidi, G., Pettenati, M. C., Melillo, P., & Iadanza, E. (2014). A machine learning system to improve heart failure patient assistance. IEEE journal of biomedical and health informatics, 18(6), 1750-1756.

[8] Zhang, R., Ma, S., Shanahan, L., Munroe, J., Horn, S., & Speedie, S. (2017, November). Automatic methods to extract New York heart association classification from clinical notes. In 2017 ieee international conference on bioinformatics and biomedicine (bibm) (pp. 1296-1299). IEEE.

[9] Parthiban, G., & Srivatsa, S. K. (2012). Applying machine learning methods in diagnosing heart disease for diabetic patients. International Journal of Applied Information Systems, 3(7), 25-30.

