IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Machine Learning Approaches to Heart Attack Risk Detection and Classification

[1]Talasila Dheeraj, [2]Skanda P R
[1]UG Scholar, [2]UG Scholar
[1]Computer Science and Engineering,
[1]R.V College of Engineering, Bangalore, India

Abstract: Cardiovascular diseases, including heart attacks, remain one of the leading causes of mortality worldwide. Early detection and accurate classification of individuals at risk of experiencing a heart attack are crucial for taking preventive measures. In this research paper, we explore various machine-learning algorithms for heart attack prediction and classification. Leveraging a dataset comprising diverse parameters, we utilize various machine-learning techniques. Our study aims to develop an efficient predictive model capable of identifying individuals susceptible to heart attacks and effectively classifying them. Through comprehensive experimentation and evaluation, we assess the performance of these models, thereby contributing to the advancement of cardiovascular health management.

IndexTerms **– Cardiovascular diseases, Classification model, Deep learning, Heart attack prediction, Machine learning**

INTRODUCTION

Cardiovascular diseases (CVDs) represent a significant global health challenge, with heart attacks being a leading cause of mortality. Timely identification and classification of individuals at risk of experiencing a heart attack are critical for implementing preventive measures and improving outcomes. Traditional risk assessment methods often rely on clinical guidelines and risk scoring systems, which may have limitations in accurately predicting risks. In recent years, the emergence of machine learning (ML) techniques has provided new opportunities to enhance heart attack prediction and classification.

This research paper uses various ML algorithms for heart attack detection and classification. By leveraging large datasets containing diverse parameters, ML models can learn complex patterns and relationships to predict individualized heart attack risks. The objective of this study is to develop predictive models capable of accurately identifying individuals at risk of experiencing a heart attack. Achieving this can facilitate targeted interventions and personalized healthcare strategies, thus improving outcomes.

We then present our methodology for data collection, preprocessing, feature selection, model development, and evaluation. The performance of different ML algorithms is systematically assessed using appropriate metrics such as accuracy, efficiency, F1 score, etc. This contributes to the growing knowledge in cardiovascular health management by demonstrating ML approaches in heart attack prediction and risk classification. The findings of this study have implications for public health interventions.

NEED OF THE STUDY

The increasing cases of cardiovascular diseases, especially heart attacks, call for the need to implement advanced methodologies in risk prediction and classification. Existing conventional approaches often lack the granularity and adaptability required to accurately discern individualized risk profiles, leading to suboptimal preventive strategies and healthcare outcomes. In this regard, the application of machine learning (ML) techniques offers an opportunity to leverage vast and diverse datasets, enabling the development of predictive models that can discern intricate patterns and interrelationships among various parameters. By harnessing ML algorithms, we can pave the way for personalized risk assessments, facilitating early interventions and tailored preventive measures. This research endeavor not only holds promise for improving outcomes but also has to reduce global cardiovascular diseases.

RESEARCH METHODOLOGY

The methodological approach adopted in this study is:

3.1 Data and Sources of Data

The research begins with the acquirement of an extensive open-source dataset sourced from Kaggle, a renowned platform for datasets. This dataset encompasses a diverse array of parameters pertinent to cardiovascular health, The utilization of such a

comprehensive dataset from Kaggle ensures the inclusion of varied features essential for robust heart attack prediction and risk classification models.

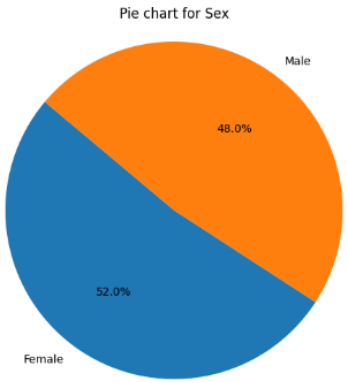
3.2 Data Preprocessing

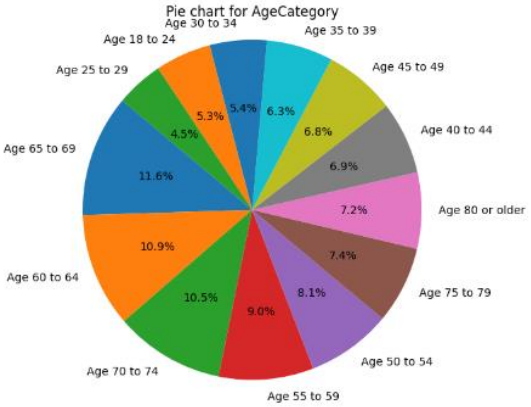
The dataset consisted of over 246,022 rows and 40 columns consisting about 98,40,880 cells of data. The columns of the dataset included very crucial clinical parameters such as State, Sex, General Health, Physical Health Days, Mental Health Days, Last Checkup Time, Physical Activities, Sleep Hours, Removed Teeth, Had Heart Attack, Had Angina, Had Stroke, Had Asthma, Had COPD, Had Depressive Disorder, Had Kidney Disease, Had Arthritis, Had Diabetes, Deaf Or Hard Of Hearing, Blind Or Vision Difficulty, Difficulty Concentrating, Difficulty Walking, Difficulty Dressing Bathing, Difficulty Errands, Smoker Status, Cigarette Usage, Chest Scan, Race Ethnicity Category, Age Category, Height In Meters, Weight In Kilograms, BMI, Alcohol Drinkers, HIV Testing, FluVaxLast12, Pneumo Vax Ever, TetanusLast10Tdap, High Risk Last Year and Covid Positive.

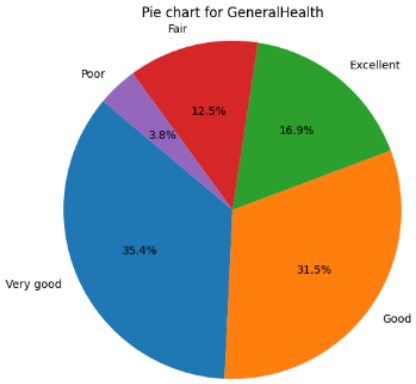
All the unique values of every column were converted to integers. For example, Former smoker: 1, Never smoked: 0, Current smoker - now smokes every day: 3 and Current smoker - now smokes some days: 2. Here the numbers are given based on severity and in some cases just for the variety (classification). This preprocessed dataset is free of any null/ missing values and is ready to be trained and deployed.

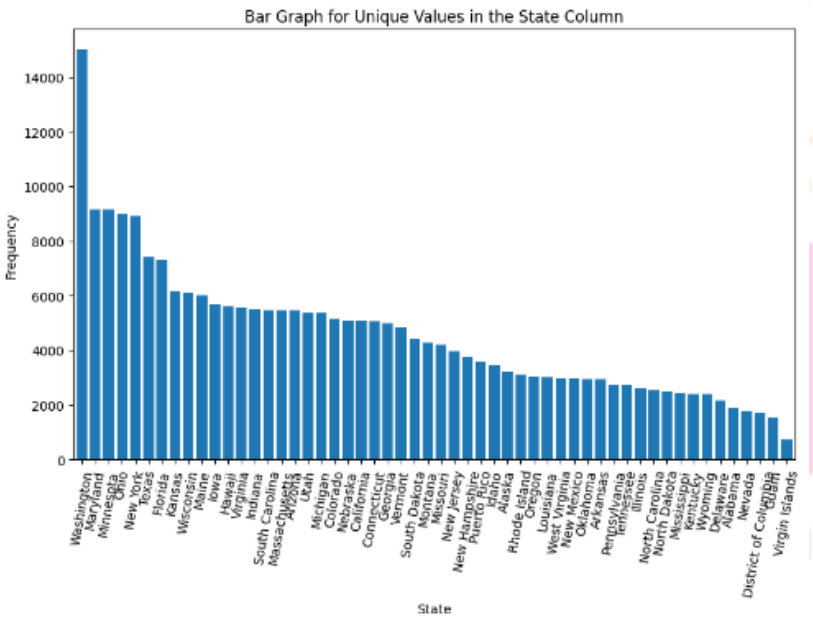
3.4 Statistical Analysis

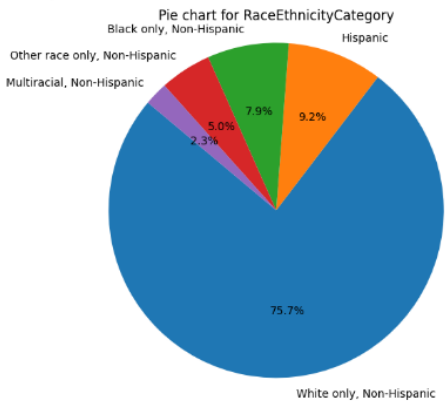
This section elaborates on the detailed statistical analysis of the dataset. They are:

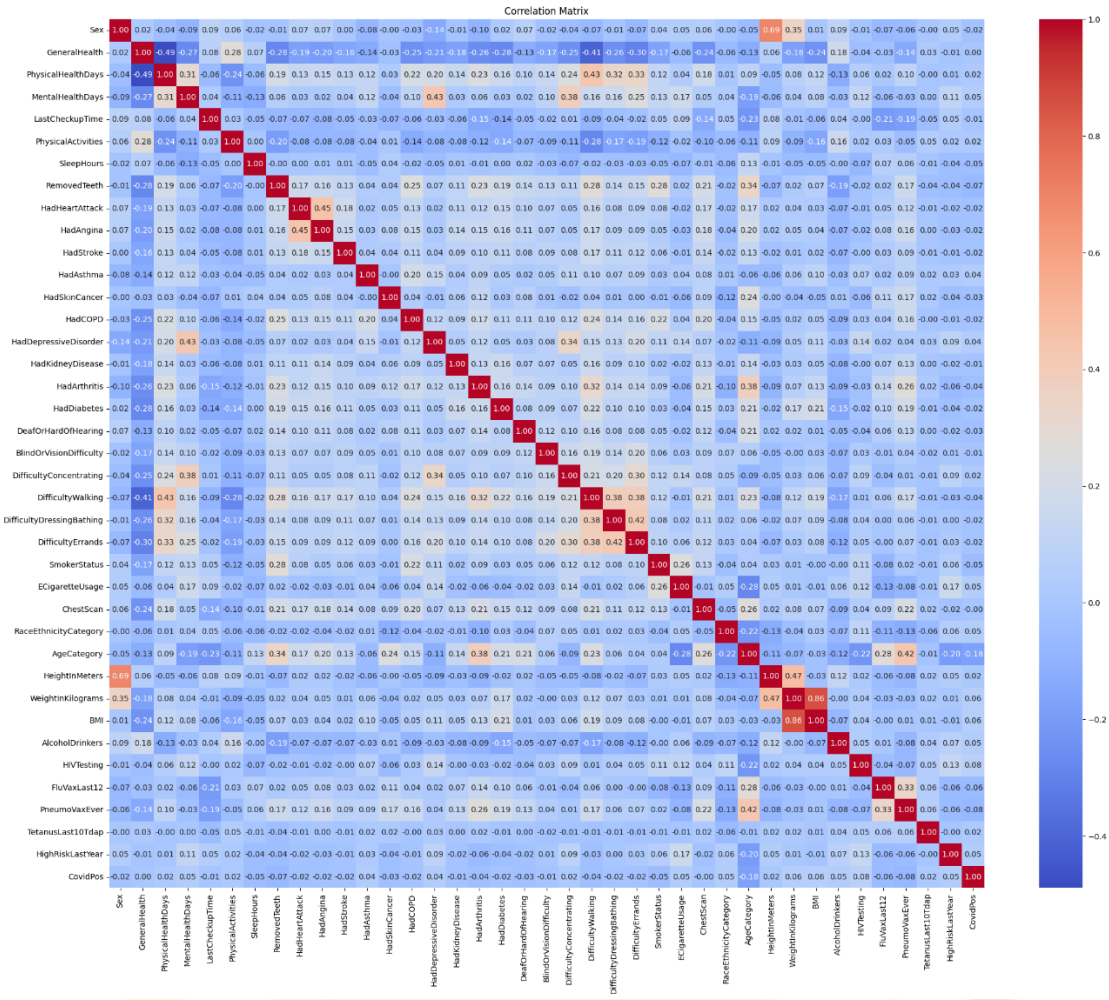












These statistical visualizations confirm the diversity and variety in the dataset. Such a diverse array of values not only highlights the heterogeneous nature of the data but also emphasizes the complexity and multidimensionality of factors influencing our classification. This depth of variation serves as a testament to the dataset's capacity to identify trends and intricate patterns, laying the foundation for robust analysis and insightful interpretations during the exploration.

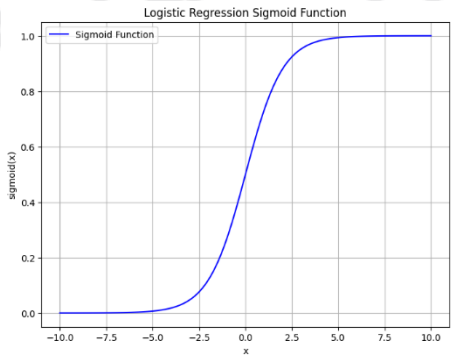
3.5 Training and Testing the Models Using Various Algorithms

The next step of the study is to build the model using various classification algorithms and compare them based on accuracy and F1 score to realize the best model for the given Diverse dataset.

3.5.1 Logistic Regression

The Dataset was trained to utilize logistic regression, which is one of the most common and basic methods used for classification and predictive analysis. The general equation and the plot of the logistic equation are given as follows:

$$P = \frac{e^{a+bX}}{1+e^{a+bX}} \qquad (3.1)$$



The confusion matrix for the classification of heart attack is

```
[[46020   553]
 [ 1978   654]]
```

3.5.2 Decision Tree Classifier

Following Logistic regression, the dataset was trained and tested using a Decision Tree classifier. The confusion matrix for the same is given as follows:

```
[[44502  2071]
 [ 1863   769]]
```

3.5.3 Random Forest Classifier

The Random Forest classifier fits n number of decision tree classifiers on various sub-samples of the dataset and uses averaging to give the final results. The confusion matrix for the same is given as follows:

```
[[46336   237]
 [ 2285   347]]
```

3.5.4 Gradient Boosting Machines (GBM)

This algorithm builds an additive model. It allows for the optimization of differentiable loss functions. In each stage n classes regression trees are fit on the negative gradient (slope) of the loss function. The confusion matrix for GBM is as follows:

```
[[46067   506]
 [ 1984   648]]
```

3.5.5 KNN Classifier

KNN is a simple and intuitive algorithm that classifies data points based on the majority class among their k-nearest neighbors in the space. The confusion matrix for the KNN classifier is shown below:

```
[[45958   615]
 [ 2080   552]]
```

3.5.6 Naïve Bayes Classifier

Naive Bayes is a probability classifier that applies Bayes' theorem with the assumption of independence between every pair of features. The confusion matrix for the same is as shown below:

```
[[38145  8428]
 [  757  1875]]
```

3.5.7 XGB Classifier

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient and flexible. The confusion matrix for the same is as follows:

```
[[46045   528]
 [ 1975   657]]
```

3.5.8 AdaBoost Classifier

AdaBoost is a popular ensemble learning algorithm used for classification tasks. It is designed to improve the performance of weak learners by combining multiple weak learners to create a strong learner. The confusion matrix is as follows:

```
[[45939   634]
 [ 1929   703]]
```

3.5.9 Deep Learning Using Neural Networks

The deep learning method aims to learn patterns in data by building complex models composed of multiple layers of interconnected neurons, also known as Neural Networks (NN)/ Artificial Neural Networks (ANNs).

```
[[46193   380]
 [ 2118   514]]
```

IV. RESULTS AND DISCUSSION

4.1 Results of Classification using various Machine Learning Algorithms

Table 4.1 Statistics of various Algorithms on the dataset

S.No	ML Algorithm	Precision	Recall	Accuracy	F1 Score
1.	Logistic Regression	0.95878994958	0.98812616752	0.94875495477	0.97323703883
2.	Decision Tree Classifier	0.95981882885	0.95553217529	0.92004877553	0.95604381407
3.	Random Forest Classifier	0.95300384607	0.99491121465	0.94874504624	0.97350673361
4.	Gradient Boosting Classifier	0.95871053672	0.98913533592	0.94939538664	0.97368532243
5.	KNN Classifier	0.95670094508	0.98679492409	0.94522914337	0.97151494011
6.	Naïve Bayes Classifier	0.98054084622	0.81903678096	0.81333197846	0.89254167884
7.	XG Boost Classifier	0.95887130362	0.98866295922	0.94913118586	0.97353926822
8.	AdaBoost Classifier	0.95970167961	0.98638696240	0.94791179758	0.97284428371
9.	Neural Networks	0.95615905280	0.99184076610	0.94923280154	0.97367314303

Based on the above table, it is evident that the Naïve Bayes classifier had high precision but was unable to maintain overall accuracy and F1 score. Ensemble methods such as the AdaBoost Classifier, XG Boost classifier, and Random Forest classifier performed well in all the metrics mentioned in the table. Logistic Regression, Decision trees Classifier, and KNN were also demonstrated very well. Neural Networks also gave competitive results, showing their ability to learn intricate patterns from complex datasets. However, the Gradient Boosting algorithm stands out best when compared to other ML models in terms of both Accuracy and F1 Score.

These findings highlight the importance of employing advanced machine learning techniques for heart attack prediction, with each algorithm contributing unique strengths to the overall predictive framework. Further exploration and refinement of these models may offer enhanced accuracy and reliability in identifying individuals at risk of experiencing a heart attack, thereby facilitating timely intervention and preventive measures.

REFERENCES

[1] J. N, D. P, M. E, R. Santhosh, R. Reshma, and D. Selvapandian, "Heart Attack Prediction using Machine Learning," 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2022, pp. 854-860, doi: 10.1109/ICIRCA54612.2022.9985736

[2] Mohamed Wed Eladham, Ali Bou Nassif, and Mohammad AlShabi "Heart attack prediction using machine learning", Proc. SPIE 12548, Smart Biomedical and Physiological Sensor Technology XX, 125480F (14 June 2023); https://doi.org/10.1117/12.2664047

[3] Bhatt, C.M.; Patel, P.; Ghetia, T.; Mazzeo, P.L. Effective Heart Disease Prediction Using Machine Learning Techniques. Algorithms 2023, 16, 88. https://doi.org/10.3390/a16020088