Heart Disease Prediction using Machine Learning Algorithms

[1]Ch.Varshith, [2]H. Varshith, [3]D. Varshitha, [4]M. Varshitha, [5]T. Varshitha, [6]G. Varun, [7]Dr. K. Satish

[123456]Student, [7] Professor

School of Engineering – IIIrd year AI&ML-ZETA

[1]2111cs020620@mallareddyuniversity.ac.in [2] 2111cs020621@mallareddyuniversity.ac.in
[3]2111cs020622@mallareddyuniversity.ac.in [4] 2111cs020623@mallareddyuniversity.ac.in
[5]2111cs020624@mallareddyuniversity.ac.in [6] 2111cs020625@mallareddyuniversity.ac.in

Department of Artificial Intelligence & Machine Learning

Malla Reddy University, Kompally, Hyderabad, India

Abstract

Presently, health conditions are on the rise primarily due to lifestyle factors and genetic predispositions. Among these, heart disease has notably surged, posing a significant risk to people's lives. Each person possesses unique benchmarks for vital health indicators like blood pressure, cholesterol levels, and pulse rates. However, according to established medical standards, normal values typically fall within certain ranges: Blood pressure ideally registers at 120/90, cholesterol between 100-129 mg/dL, pulse rate around 72, fasting blood sugar level at 100 mg/dL, heart rate ranging from 60-100 beats per minute (bpm), normal ECG readings, and major vessel widths spanning from 25 mm (1 inch) in the aorta to a mere 8 µm in capillaries.

This study explores various classification techniques employed to predict the risk levels of individuals based on parameters such as age, gender, blood pressure, cholesterol levels, and pulse rate. The "Disease Prediction" system relies on predictive modeling to anticipate a user's disease risk by analyzing symptoms provided as input. By evaluating user-input symptoms, the system computes the probability of specific diseases as an output. Disease prediction involves the implementation of five techniques: Naïve Bayes, KNN, Decision Tree, Linear Regression, and Random Forest Algorithms. These methods gauge the likelihood of an individual developing a particular ailment. Consequently, the collective average prediction accuracy probability stands at an impressive 83%.

Keywords:

Naïve Bayes, KNN, Decision Tree, Linear Regression & Random Forest.

Introduction

In our daily lives, numerous factors impact the health of the human heart, and the identification of new heart conditions is rapidly increasing. The heart, an indispensable organ responsible for circulating blood throughout the body, is crucial for overall health and well-being. Its condition is greatly influenced by an individual's professional and personal behaviors, shaped by life experiences and sometimes influenced by genetic predispositions, wherein certain heart diseases are inherited across generations.

According to the World Health Organization (WHO), over 12 million deaths occur annually worldwide due to various heart diseases, collectively known as cardiovascular diseases. This umbrella term encompasses a range of conditions specifically affecting the heart and arteries. Alarmingly, even individuals in their 20s and 30s are increasingly falling victim to heart diseases. This surge among younger demographics is attributed to unhealthy eating habits, sleep deprivation, stress, depression, and multiple factors including obesity, poor dietary choices, family medical history, high blood pressure, elevated cholesterol levels, sedentary lifestyles, smoking, and hypertension.

The healthcare sector stands as one of the most crucial and economically substantial industries in the 21st century. Addressing affordability and ensuring quality in healthcare services requires extensive statistical analyses, especially in combating the rising incidence of chronic diseases. Advanced data-driven intelligent technologies play a pivotal role in disease diagnosis, detection, treatment, and research.

This paper focuses on proposing a model for predicting and diagnosing cardiovascular diseases by integrating electrocardiogram (ECG) analysis and symptom-based detection. The aim is to develop a robust and reliable research tool that continually evolves through further research. The paper delves into classical methods and algorithms employed in cardiovascular disease (CVD) prediction, highlighting gradual advancements and comparing the performance of existing systems. It presents an improved multi-module system aimed at enhancing accuracy and feasibility.

The model's implementation involves utilizing datasets from repositories such as UCI and PhysioNet, where modifications in data format were made, particularly in ECG reports, to optimize the convolutional neural network used in our research. For risk prediction, specific attributes were selected for training and implementing a multi-layered neural network developed by our team.

The paper concludes by outlining potential avenues for further research and advancements in this field.

Literature Review

Ordonez proposed a method for predicting heart disease by utilizing fundamental patient attributes, totaling 13, including sex, blood pressure, cholesterol, among others. To enhance their research dataset, they introduced two additional attributes: fat and smoking behavior. Their system aimed to predict the likelihood of an individual being affected by heart disease. They employed data mining classification algorithms such as Decision Tree, Naive Bayes, and Neural Network on a Heart disease database to make predictions. The study analyzed the outcomes derived from these algorithms.

Data mining plays a pivotal role in forecasting heart disease. In the realm of medical data mining, there exists substantial potential for uncovering hidden patterns crucial for clinical diagnosis across various disease datasets. Multiple data mining techniques, including Naive Bayes, Decision Tree, neural networks, kernel density estimation, bagging algorithms, and support vector machines, have been employed for heart disease diagnosis, each demonstrating varying levels of accuracy. Notably, Naive Bayes stands out as a successful classification technique in predicting heart disease among patients.

Peter et al. discussed a novel feature selection method algorithm that combines the CFS (Correlation-based Feature Selection) and Bayes theorem, denoted as CFS+Filter Subset Eval. This hybrid method was evaluated and achieved an accuracy rate of 85.5%, showcasing promising results in enhancing the accuracy of heart disease prediction models.

Methodology

The research methodology employed in this paper involved a comprehensive literature review of existing studies and research in the realm of Heart Disease Prediction. The focus was on identifying accurate Machine Learning Algorithms for precise predictions. The data sources for this review included scholarly articles, research papers, and relevant books that shed light on Heart Disease Prediction and effective Machine Learning Algorithms.

The collected data from these sources underwent analysis to discern the key features crucial for accurate Heart Disease Prediction using Machine Learning Algorithms. The primary objective of this project was to predict occurrences of heart disease with the utmost accuracy. To achieve this goal, multiple classification algorithms were tested.

Firstly, a general function utilizing the SciKit Learn library was formulated to facilitate the training of models. Accuracy assessments were conducted on both training and test sets to determine potential overfitting or underfitting issues, known as the bias/variance tradeoff.

A. Logistic Regression:
This classification algorithm assigns observations to discrete classes. Unlike linear regression, which produces continuous values, logistic regression employs the logistic sigmoid function to transform outputs into probability values, applicable to two or more discrete classes. The accuracy score obtained using Logistic Regression was 85.25%.

B. Random Forest:
A supervised learning algorithm suitable for both classification and regression problems. In this study, Random Forest was employed solely for classification purposes. The accuracy score achieved using Random Forest was 86.9%.

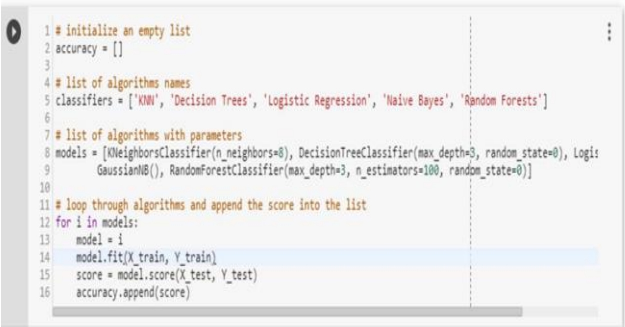
C. Naïve Bayes:
Utilizes Bayes' Theorem to calculate the probability of a hypothesis given the data. The accuracy score of Naïve Bayes in this research was 85.25%.

D. K-Nearest Neighbor (KNN):
This model predicts classes by calculating distances between test data and training data points, then selecting the most frequent class among the nearest neighbors. The accuracy score obtained using KNN was 68.25%.

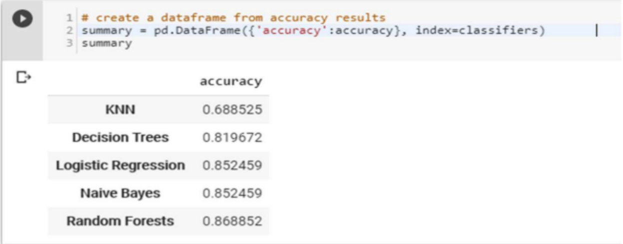
E. Decision Trees:
Decision Trees were constructed by placing the best attribute at the root and splitting the training set into subsets based on attribute values. The model achieved an accuracy score of 81.96%.

Each algorithm was assessed for its predictive capability in determining occurrences of heart disease based on the features provided, contributing unique insights into the accuracy and performance of these Machine Learning models in this domain.

Results

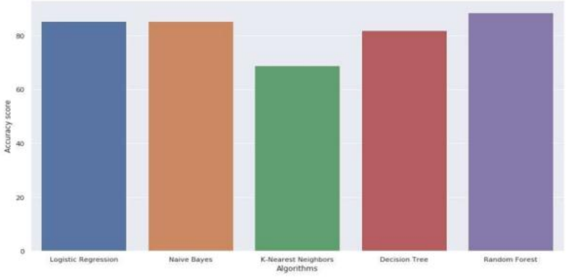


```
# initialize an empty list
accuracy = []

# list of algorithms names
classifiers = ['KNN', 'Decision Trees', 'Logistic Regression', 'Naive Bayes', 'Random Forests']

# list of algorithms with parameters
models = [KNeighborsClassifier(n_neighbors=8), DecisionTreeClassifier(max_depth=3, random_state=0), Logis
        GaussianNB(), RandomForestClassifier(max_depth=3, n_estimators=100, random_state=0)]

# loop through algorithms and append the score into the list
for i in models:
    model = i
    model.fit(X_train, Y_train)
    score = model.score(X_test, Y_test)
    accuracy.append(score)
```



```
# create a dataframe from accuracy results
summary = pd.DataFrame({'accuracy':accuracy}, index=classifiers)
summary
```

	accuracy
KNN	0.688525
Decision Trees	0.819672
Logistic Regression	0.852459
Naive Bayes	0.852459
Random Forests	0.868852



Discussion

Our project's main aim is to accurately predict the presence of heart disease using a minimal number of tests and attributes, ensuring efficient and swift predictions. We've specifically focused on fourteen key attributes that serve as the core basis for tests, aiming to yield highly accurate results. While additional attributes could be included, our emphasis is on predicting heart disease risk within a specific age range using a streamlined approach.

To achieve this objective, we implemented five data mining classification techniques: K-Nearest Neighbor, Naive Bayes, Decision Tree, Random Forest, and Logistic Regression. Through our analysis, we found that Random Forest outperforms other techniques in terms of accuracy. It emerged as the most effective model for predicting patients with heart disease.

This project offers diverse solutions, each with its unique strengths. These models vary in their ease of interpretation, accessibility to detailed information, and accuracy in predicting heart disease. By exploring these various techniques, we've aimed to provide effective methods for predicting heart disease while considering factors like simplicity, comprehensibility, and accuracy.

Drawback

The algorithms utilized in our project do not achieve 100% accuracy, hence rendering the predictions less than fully reliable. As a result, there might be slight discrepancies between clinical diagnosis and our project's predictions. Medical diagnosis is a crucial task that demands precision and efficiency. While automation of this process could offer substantial benefits, it's important to note that clinical decisions often rely on a doctor's intuition and experience rather than solely on data obtained from datasets rich in knowledge.

Conclusion

There's ample room for further enhancement and expansion of this project. For instance, it can broaden its scope by including additional medical attributes beyond the initial 14 attributes used. Furthermore, exploring diverse data mining techniques like Time Series analysis, Clustering, and Association Rules could enrich the predictive capability. Instead of solely relying on categorical data, incorporating continuous data could provide more comprehensive insights.

Another promising avenue for improvement involves delving into Text Mining to extract valuable information from the vast pool of unstructured data available. This project, which utilizes data mining techniques such as logistic regression, KNN, Naive Bayes, Decision Tree, and Random Forest, has demonstrated the effectiveness of Random Forest in yielding superior results. Its outcomes prove beneficial for domain experts and medical professionals, aiding in planning for improved and earlier diagnosis for patients. Importantly, this system exhibits strong performance even without the need for continuous retraining.

References

Referred the below in internet:

[1] M. K. Awang and F. Siraj, “Utilization of an artificial neural network in the prediction of heart disease,” Int. J. Bio-Science Bio-Technology, vol. 5, no. 4, pp. 159–165, 2013.

[2] P. Selvakumar and S. P. Rajagopalan, “A survey on neural network models for heart disease prediction,” J. Theor. Appl. Inf. Technol., vol. 67, no. 2, pp. 485–497, 2014.

[3] A. Methaila, P. Kansal, H. Arya, and P. Kumar, “Early Heart Disease Prediction Using Data Mining Techniques,” vol. 6956, no. October, pp. 53–59, 2014.

[4] I. S. F. Dessai, “Intelligent Heart Disease Prediction System Using Probabilistic Neural Network,” no. 5, pp. 38–44, 2013.

[5] T. Karayilan and Ö. Kiliç, “Prediction of Heart disease using neural network,” in 2nd International Conference on Computer Science and Engineering, UBMK 2017, 2017, pp. 719–723.

[6] M. Mardiyono, R. Suryanita, and A. Adnan, “Intelligent Monitoring System on Prediction of Building Damage Index using Neural-Network,” TELKOMNIKA (Telecommunication Comput. Electron. Control., vol. 10, no. 1, p. 155, 2015.

[7] N. Guru, A. Dahiya, and N. Rajpal, “Decision support system for heart disease diagnosis using Neural Network,” 2007.

[8] J. S. Singh Navdeep, “No Title,” Intell. Hear. Dis. Predict. Syst. using CANFIS Genet. Algorithm, Int. J. Biol. Med. Sci., no. International Journal of Advance Research, Ideas and Innovations in Technology© 2018, www.IJARIIT.comAll Rights, p. Reserved Page | 987, 2008.