International Journal of Computer Sciences and Engineering
Vol.**11**, Special Issue.**1**, pp.**296-300**, November **2023**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Paper

Heart Diseases Prediction Model Using Density Based Clustering

Sayan Chakraborty[1*], Trisha Mondal[2], Sayantan Maity[3], Saikat Pahari[4]

[1]Dept. of Computer Science and Engineering, OmDayal Group of Institutions, Howrah, India

**Corresponding Author: sayanchakraborty710@gmail.com*

Abstract: The condition that is most prevalent nowadays is heart disease, that may be successfully treated if caught and treated at an early enough stage. Heart disease diagnosis requires extreme caution since the procedure might be derailed by human mistake. Machine learning techniques were widely popular in many walks of life, but they rose to prominence in the field of heart disease forecasting. Many biological characteristics included in cardiac patient datasets have little bearing on diagnosis. Prediction accuracy for cardiac patients may be improved while computational complexity is reduced by eliminating irrelevant elements from the available data-set. This technique provides a density-based unsupervised method for identifying cardiac anomalies. The filter-based feature selection strategy is used to begin the process of narrowing down the data collection to its most fundamental characteristics. In order to improve the clustering effectiveness of healthy cases and to detect aberrant examples like cardiac patients, a new method for clustering with adaptive variables called Density Based Clustering has been applied. The DBSCAN method, that generates density-based clusters, is intended to solve these problems; though, the best way to choose an epsilon value and a minimum value is still up for debate. These two factors are used in the suggested strategy to achieve high diagnostic accuracy in patients with cardiac conditions.

Keywords: Heart Diseases, Diseases Prediction Model, Outlier Data, Machine Learning

1. Introduction

The most frequent health problem in the contemporary world is heart disease. Twelve million people lose their lives every year because of heart disease. Diseases of the heart include a broad spectrum of conditions that affect the organ and capacity to circulate blood all around the human body. Correct and precise identification of cardiac disease is essential. Heart-related disorders may be foreseen that preventative steps can be done to prevent or suppress them if one takes into account the patient's history and way of life in a timely manner. This aid will be more substantial and lead to cost reduction if we use methods associated with the medical-information system.

Accurately diagnosing and forecasting heart disease requires taking a number of measures, one of the most important that is learning to recognize the key indicators to comprehend the progression of heart disease Logistic regression, support vector machines, as well as other machine learning techniques and random forest learning have shown useful as decision support systems for individualized cardiovascular disease forecasting. Many articles have been written on this topic. A fusion model was investigated by the researchers, who used many ML models to predict heart disease. The authors have shown the best model among them in their research [3]. To enhance the diagnostic procedure, according to Repaka *et al.* (2019) the authors used a naïve Bayesian supervised learning method to predict heart diseases. The results demonstrated that the model outperformed its competitors [2]. However, in machine learning, anomalies may occur and hinder the accuracy of the prediction model. Density-based clustering of space with noise (DBSCAN) has been shown in previous research to be an effective tool for recognizing and minimizing anomalous information.

Applying a density-based unsupervised methodology, this technology may identify irregularities in cardiac patients. Selecting the appropriate number of clusters as well as determining the most important node are two of the main challenges of partition clustering approaches. The main issue is the high error rate associated with clustering. DBSCAN is a clustering approach that uses the distance between samples and the amount of samples in a certain radius to solve these issues. Therefore, DBSCAN does not need an estimate of the number of clusters or an establishment of an arbitrary starting point. The density of the data is used to determine clusters, and when there are several samples that are quite similar, a cluster is formed. However, there is still an issue with this approach, that of course is figuring out the number of samples that should be considered to be in the neighborhood and what the criterion for determining the shortest distance between samples should be. Our solution uses the DBSCAN clustering approach to locate and forecast patients with cardiac disease.

2. Related Work

Using the Heart Diseases Dataset, several researchers have attempted to develop methods for illness prediction.

Following are some examples of Machine Learning algorithms that have been used to achieve various levels of accuracy. In order to solve this problem, Classification models are very useful as they are supervised and can classify the diseased patients accurately [1]. According to Indrakumari et al. (2020) conducted a comparison study by determining key features and using a variety of ML algorithms. They analysed their data of heart diseases using exploratory data analysis [5]. The researchers introduced a method that combines Mean Fisher Score Selection of Features Algorithm for Support Vector Machine. Classification Features with a higher Fisher score than the average are chosen [4]. SVM was trained and validated using the chosen subset of features to determine MCC. According to the results, the precision, specificity, and sensitivity may reach 81.19 percent, 72.92 percent, and 88.68 percent, respectively, when using a combination of the Fisher score selection of features algorithm (FSFSA) and the Support Vector Machine (SVM) algorithm [7].

According to Fitriyani et al. (2020) to enhance the functionality of the HDPM developed a Hybrid Random Forest with a Linear Model (HRFLM). Achieving 88.4% accuracy, 90.1% precision, 92.8% sensitivity, 90% f-measure, and 82.6% specificity was shown to be possible using the suggested strategy [13]. A genetic conduct with the goal of selecting the most relevant characteristics for cardiac illnesses was suggested by the researchers for use in support vector machine optimization [11]. According to Asadi et al. (2021) discovered a M.I. framework Random_Forest foundations of Machine Learning Architecture (MLA) went into effect. Random Forest was utilized for illness prediction once relevant characteristics were identified using Factor Analysis of Mixed Data. The experimental findings demonstrated that the suggested strategy achieved higher levels of accuracy (93.44%), sensitivity (89.28%), and specificity (96.96%) than competing models and earlier studies' results [6]. Predicting heart disease using a machine learning (ML) method that combines decision trees and random forests. Here, just the aforementioned methods have been applied to the datasets in question [10].

Several ML techniques and deep learning on a dataset of cardiac illnesses with the intention of comparing and analyzing the outcomes. To cut down on labelling expenses, El-Hasnomy and colleagues [8] implemented five multilabel active learning algorithms and utilized them to identify the most relevant data to query the labels. According to Richardson et al. (2020) to predict heart diseases they have used multivariable meridian randomization to get better accuracy [9]. Artificial Intelligence based prediction using Random forest classifier and naive bayes algorithms is better in case of heart diseases prediction [15]. None of this prior research, including the instance of a heart disease dataset, utilized an outlier verification and data balancing strategy to enhance the precision of the classification. In order to predict the anomalies in the dataset of heart diseases then optimization method is very essential [18].

3. Reference Model

Our proposed algorithm Heart diseases prediction using density based clustering, mainly based on outlier detection of patients and removing that to find efficient clusters. The objective of the algorithm is mainly to find the outliers, meaning the patients who are not heart diseases affected and clusters the affected patients in different groups categorized by density [16]. The proposed approach employs the clustering technology developed by Density Based Spatial Clustering for Applications with Noise (DBSCAN). Martin Ester and others first used it in 1996. This method is effective for clusters that are located in both dense and less dense locations. It does this through clustering together similar data points. The necessity of this approach is to remove the outliers from the dataset [14]. Through analysing the density of the aforementioned data points, this clustering approach is able to generate groups from massive datasets [17]. DBSCAN clustering's primary use is in locating anomalies. In contrast to K-Means, where we must provide the total amount of cluster focus, this approach does not need us to know that number in preparation.

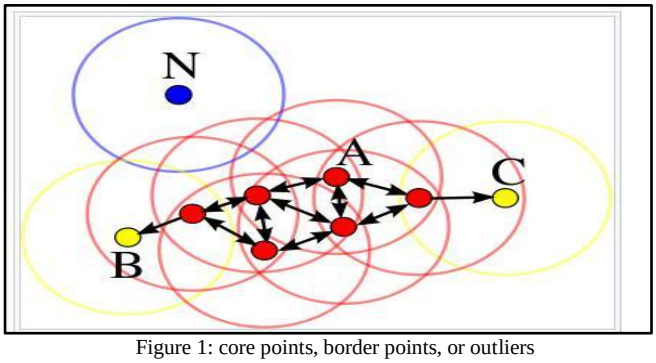


Figure 1: core points, border points, or outliers

There are two parameters of DBSCAN are eps,The metric used to denote proximity to certain areas. If the distance between two locations is less than or equal to epsilon, we call them neighbours. minPts, Few data points to define a cluster. Based on these two parameters, points are divided into core points, border points, or outliers. Core Point refers to a point with fewer than minPts points within its immediate surrounding area. The point which is near to the core point and less minPts, number of points in its area is called Border Point. The point which does not come under core point and not near from any core point is called Outlier.

4. Flowchart

Flowchart has been a powerful tool, a critical asset that serves as a guiding light in navigating and perceiving the intricate labyrinth of the program's code. It is not just a diagram but a visual structure that organizes the systematic approach underlying the data processing, analysis, and visualization efforts. By merit of its clarity and precision, it bequeaths to all stakeholders a lucid and profound understanding of the code's inner workings, lay down its functionality and potential. It offers a key to unlock the treasure trove of insights that lie dormant within, poised to revolutionize the realm of population health management.

The utilization of flowcharts within this program indicates the dedication to a methodical and structured approach to deal with the complexities of data processing, analysis, and visualization. It is similar to the blueprint of an architectural masterpiece, meticulously delineating each step, each operation, and each decision point in the journey of the data. It serves as a guiding star, allowing stakeholders to traverse this intricate landscape with confidence and clarity. In doing so, it empowers them to not only understand but to wield the code effectively, like a masterful conductor leading an orchestra to produce harmonious and enlightening melodies. The figure 2 shows the flowchart that visualizes steps of the algorithm.

In the grand tapestry of our program's code, the flowchart is the dyer's loom, fastidiously crafting each elaborate pattern and connection. It is the cartographer's map, charting the course through uncharted territory. It is the composer's score, orchestrating the symphony of data processing, analysis, and visualization. It is the storyteller's narrative, unfolding the tale of our code's journey from raw data to meaningful insights. It is, above all, the torchbearer of knowledge and innovation, guiding stakeholders toward a brighter future in population health management.

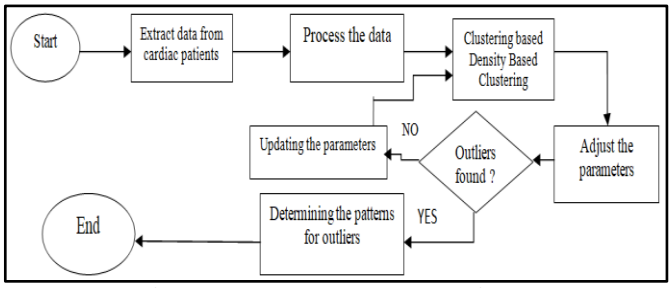


Figure 2: flowchart diagram to show the steps of our algorithm

5. Proposed Algorithm

The steps provided outline a sequence of actions to create a data visualization using Python, specifically focusing on importing libraries, processing data from a CSV file, and generating various types of plots. These steps utilize the pandas and matplotlib libraries, two powerful tools in the Python data science ecosystem.

- Step 1. START-This step represents the beginning of the process, which is the initiation of the project of creating a data visualization. It is necessary for the starting point of your code.
- Step 2. Import matplotlib and pandas-In this step, two important libraries must be imported, matplotlib and pandas. Matplotlib is a wide data visualization library, whereas pandas is a flexible data manipulation library. These libraries will be used to handle and visualize your data.
- 3. Use pandas to specify the csv file location containing the survey data- In this step use pandas to specify the location of the CSV file that contains your survey data. Pandas supplies functions to read data from various file formats, including CSV.
- Step 4. Initialization of X and Y axis of the graph using csv data: This step includes loading the data from the CSV file into variables that will represent axes of your graph. This data will be used to create the plots.
- Step 5. Add the title of the graph, the X axis label and the Y axis label: In this step, set up the common elements of the graph. You add a title to elaborate the purpose of the graph, labels for the X and Y axes to discuss context.
- Step 6. Use of matplotlib to plot the scatter points using the data from the csv file: Matplotlib provides a variety of plot types, and in this step, you use it to create a scatter plot. A scatter plot is useful for visualizing individual data points, such as survey responses, to understand the distribution or patterns within the data.
- Step 7. Clustering based on Density based method: This step is responsible for clustering the dataset using the Density based clustering method.
- Step 8. Adjust parameters: Here, the parameters of the density clustering method must be adjusted which are epsilon and minimum points.
- Step 9. Updating parameters: The step will be proceeded if outliers are not found.
- Step 10. Determining patterns for outliers: After finding the outliers now the outliers must be visualized in different colors to separate the clusters and the outliers.
- Step 11. STOP: This step signifies the end of the process. Your data visualization is complete, and you can now analyze, share, or save the results as needed.

In summary, these steps provide a structured framework for creating data visualizations in Python, specifically utilizing the pandas and matplotlib libraries. The process includes importing data, preparing it, customizing the graph, plotting the data, and implementing the density clustering method and context. The final result is a powerful tool for understanding and conveying insights from your data.

6. Results and Discussion

The results of this research are visualised in this chapter . This chapter is responsible for presenting the result as tables, graph plots. In this chapter a table is shown that describes the columns of the dataset, some graph plot that shows the best epsilon value for the research and also shown in table presentation.

Upon the completion of data collection, a rigorous analysis ensued. This included sifting through the gathered information. One of the important outcomes of this analysis was the computation of the data. These data has supplied critical insights into the prevalence of the three targeted health condition, heart disease.

In order to assure the accessibility and utility of these calculated data, they were thoughtfully organized and documented in a proper column. This step was important, as it has served as a reference point for future research and analysis and also enabled a wide understanding of the health landscape within each district. This methodical approach to data collection, analysis, and documentation laid the foundation for evidence-based decision-making and targeted

interventions to improve public health. After meticulously collecting and organizing the data, it has proceeded to employ the code to create graphical representations of the prevalence of heart diseases. These visualizations offer a clear and concise overview of the data's insights.

Table 1: Descriptions of the attributes

"Sl. No.	Attribute Description	Distinct values of Attribute
1.	*Age*- represents the age of the patient.	Multiple values between 29 and 71
2.	*Sex*- describe the gender of the patient(0=female,1=Male).	0,1
3.	*Chest pain type*- represents the severity of chest pain a patient is suffering.	0,1,2,3
4.	*Resting BP*- This represents the patient's blood pressure.	Multiple values between 94 and 200
5.	*Cholesterol*- It shows the cholesterol level of the patient.	Multiple values between 126 and 564
6.	*Fasting Blood Sugar*- It represents the fasting blood sugar in the patient.	0,1
7.	*Resting ECG*- It shows the result of ECG.	0,1,2
8.	*Max Heart Rate*-Shows the maximum heart rate of the patient.	Multiple values between 71 and 202
9.	*Exercise angina*-used to identify if there is an exercise including angina. If yes=1 or-else no=0.	0,1
10.	*OldPeak*-describes a patient's depression level.	Multiple values between 0 and 6.2
11.	*Slope*-describes patient condition during condition during peak exercise. It is divided into three segments(Up-sloping, Flat, Down-Sloping).	1,2,3"

The generated graphs enable us to discern patterns of the clusters for distinct cluster patterns for different values of epsilon and minimum values. By examining these graphs, it has been identified regions or districts with higher or lower incidences of this health condition. This information is invaluable for healthcare planning and resource allocation, as it helps target interventions and support to areas where they are most needed. The graphical representations serve as a tool for communication and decision-making. They offer a visual narrative of heart disease, making it easier for healthcare professionals to comprehend the data at a glance. This aids in the formulation of evidence-based strategies to address health disparities and improve the overall well-being of the population.

In summary, the utilization of the code to create these graphs marks a significant step in transforming raw data into actionable insights. It empowers us to better understand heart disease, guiding toward informed decisions and effective healthcare management.

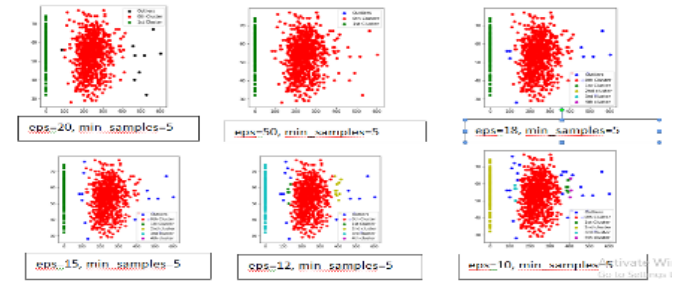


Figure 3: Cluster Visualization

It is visualised at the distribution of distinct epsilon values after analyzing the findings of the graph plotting. According to the color-coded graphs, the outliers are represented by blue, the cluster1 by red, and the cluster 2 by yellow. In order to compare the average number of affected districts, another table has also been created, which sheds light on the overall impact of the epsilon values to get optimal values.

Table 2: Eps and corresponding outlier

Min_samples	Eps[Epsilon]	Number of Outliers
5	50	0
5	20	9
5	15	16
5	12	18

7. Conclusion

This chapter is responsible for concluding this research and discussing the future scope of this research. In this research the dataset has been collected through kaggle.com. The dataset has described the patient and their health details. In this research the scatter graph is visualized that has come after implementing the proposed theory that is Density based clustering which creates clusters on the basis of dense regions. The epsilon value that is responsible to create the clusters and differentiate them is taken as 20 with which the outliers were better differentiated. The accuracy of the proposed model is 76.53 percent.

This part concludes by stating that spatial clustering based on density for applications with noise (DBSCAN) is the best clustering strategy for handling large datasets. This method clusters the data-points as density and separates the noise or outliers. It will help clinicians to make decisions more efficiently. We will also develop a mobile app so that patients can be aware about his/her heart condition and treatment required. Furthermore the epsilon value can be more accurate after using the voronoi diagram and the voronoi circle which the researchers are planning to implement later.

Data Availability

The data for this project was obtained from kaggle.com, a trusted platform for accessing a wide range of datasets. We're grateful to the data.world community for sharing this valuable resource.

Conflict of Interest

The authors declare that there is no conflict of interest.

Funding Source

Neither it is applicable, nor any funding have been used.

Acknowledgement

We would like to express our heartfelt gratitude to everyone who contributed to the successful completion of this project "Heart Diseases Prediction using Density Based Clustering". Our sincere thanks go to our advisor for their guidance, support, and valuable insights throughout the research process. We are also deeply appreciative of the participants who willingly shared their experiences and perspectives, making this study possible. Additionally, we extend our

thanks to the authors of the research papers that informed this study. Finally, we want to acknowledge the unwavering support and encouragement from our friends and family, without whom this project would not have been accomplished.

Author's Contribution

Sayan Chakraborty was involved in retrieving data and information for this particular paper and did the coding. Trisha Mondal was involved in writing these chapters: abstract, introduction, related works and the references. Sayantan Maity has written the proposed theory, experiment method and result discussion chapters. Saikat Pahari has helped in formatting, and guiding about the information needed in the paper. Always encouraged to participate in this event. He supported and gave us ideas according to this paper.

REFERENCES

[1] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," Informatics in Medicine Unlocked, vol. 19, p. 100330, 2020. doi:10.1016/j.imu.2020.100330

[2] A. N. Repaka, S. D. Ravikanti, and R. G. Franklin, "Design and implementing heart disease prediction using naives bayesian," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019. doi:10.1109/icoei.2019.8862604

[3] A. Singh and R. Kumar, "Heart disease prediction using machine learning algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), 2020. doi:10.1109/ice348803.2020.9122958

[4] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using Machine Learning Techniques," SN Computer Science, vol. 1, no. 6, 2020. doi:10.1007/s42979-020-00365-y

[5] M. A. Khan, "An IOT framework for heart disease prediction based on MDCNN classifier," IEEE Access, vol. 8, pp. 34717–34727, 2020. doi:10.1109/access.2020.2974687

[6] M. Tarawneh and O. Embarak, "Hybrid approach for heart disease prediction using data mining techniques," Advances in Internet, Data and Web Technologies, pp. 447–454, 2019. doi:10.1007/978-3-030-12839-5_41

[7] N. Kagiyama, S. Shrestha, P. D. Farjo, and P. P. Sengupta, "Artificial Intelligence: Practical primer for clinical research in cardiovascular disease," Journal of the American Heart Association, vol. 8, no. 17, 2019. doi:10.1161/jaha.119.012788

[8] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "HDPM: An effective heart disease prediction model for a clinical decision support system," IEEE Access, vol. 8, pp. 133034–133050, 2020. doi:10.1109/access.2020.30105

[9] R. Indrakumari, T. Poongodi, and S. R. Jena, "Heart disease prediction using exploratory data analysis," Procedia Computer Science, vol. 173, pp. 130–139, 2020. doi:10.1016/j.procs.2020.06.017

[10] S. Asadi, S. Roshan, and M. W. Kattan, "Random forest swarm optimization-based for heart diseases diagnosis," Journal of Biomedical Informatics, vol. 115, p. 103690, 2021. doi:10.1016/j.jbi.2021.103690

[11] S. E. Ashri, M. M. El-Gayar, and E. M. El-Daydamony, "HDPF: Heart disease prediction framework based on hybrid classifiers and genetic algorithm," IEEE Access, vol. 9, pp. 146797–146809, 2021. doi:10.1109/access.2021.3122789

[12] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, vol. 7, pp. 81542–81554, 2019. doi:10.1109/access.2019.2923707

[13] T. G. Richardson et al., "Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable mendelian randomisation analysis," PLOS Medicine, vol. 17, no. 3, 2020. doi:10.1371/journal.pmed.1003062

[14] U. Nagavelli, D. Samanta, and P. Chakraborty, "Machine learning technology-based heart disease detection models," Journal of Healthcare Engineering, vol. 2022, pp. 1–9, 2022. doi:10.1155/2022/7351061

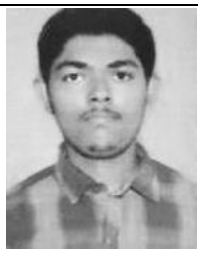
[15] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and naive bayes," The Journal of Supercomputing, vol. 77, no. 5, pp. 5198–5219, 2020. doi:10.1007/s11227-020-03481-x

[16] H. Santoso and A. Musdholifah, "Case base reasoning (CBR) and density based spatial clustering application with noise (DBSCAN)-based indexing in medical expert systems," Khazanah Informatika : Jurnal Ilmu Komputer dan Informatika, vol. 5, no. 2, pp. 169–178, 2019. doi:10.23917/khif.v5i2.8323

[17] Y. A. Nanehkaran et al., "Anomaly detection in heart disease using a density-based unsupervised approach," Wireless Communications and Mobile Computing, vol. 2022, pp. 1–14, 2022. doi:10.1155/2022/6913043

[18] S. Kannan, "Modelling an efficient clinical decision support system for heart disease prediction using learning and optimization approaches," Computer Modeling in Engineering & Sciences, vol. 131, no. 2, pp. 677–694, 2022. doi:10.32604/cmes.2022.018580

AUTHORS PROFILE



Sayan Chakraborty has pursued a Bachelor of Technology in computer science from Omdayal Group of Institutions. He is interested in researching Machine learning and other projects related to artificial Intelligence.

Trisha Mondal has pursued a Bachelor of Technology in computer science from Omdayal Group of Institutions. Her research interest is Python, Machine Learning and also on different projects based on Artificial Intelligence.

Sayantan Maity has pursued a bachelor of technology in computer science from Omdayal Group of Institutions. He is interested in researching Machine learning and other projects related to artificial Intelligence.

Saikat Pahari received Bachelor of Computer Science and Engineering in 2002 and Master of Computer Science in 2007 from university of Calcutta. He is currently working as Assistant Professor in the Department of Computer Science in Omdayal Group of Institutions , Howrah, India. He has published several research papers in reputed international journals including Thomson Reuters, Scorpus and conferences and it's also available online. His main research work focuses on Algorithms, Object Detection, Data Science, Complex Network, Big Data Analytics, Data Mining and Computational intelligence based education. He has 15 years of teaching experience and 4 years of industry experience and 6 years of research experience.

