Analysis of traditional machine learning approaches on heart attacks prediction

Micheal BERDINANTH, Samah SYED, Shudhesh VELUSAMY, Angel Deborah SUSEELAN*, Rajalakshmi SIVANAIAH

Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, India

***Corresponding Author:**

Angel Deborah SUSEELAN
angeldeborahs@ssn.edu.in

Abstract: Considering the persistent challenge of early heart attack detection in patients, despite significant advancements in medical systems, this research project is motivated by the imperative need to develop effective predictive machine learning models. The central problem addressed here in is the identification of individuals at risk of experiencing a heart attack. In response to this problem, two distinct models have been devised and meticulously evaluated, namely decision trees and logistic regression, each designed to fulfil the primary objective of this research. Through a rigorous analysis and thorough evaluation of the results, we have scrutinised the performance of these models. The comparison between decision trees and logistic regression provides valuable insights into their efficacy in predicting heart attacks. The culmination of this endeavor not only contributes to the growing body of knowledge in heart attack prediction and provides healthcare professionals with powerful tools for early diagnosis, potentially saving lives and improving patient outcomes.

Keywords: Machine Learning, Heart Disease, Classification, Feature Selection, Prediction.

1. Introduction

Heart disease and stroke continue to be leading causes of mortality worldwide, claiming the lives of over 17.8 million individuals each year, according to the World Health Organization (WHO). Despite the vast amounts of healthcare data generated daily, the full potential of this information remains untapped in transforming patient outcomes. Cardiovascular diseases, including coronary artery disease and myocarditis, have a heavy toll, with 80% of all deaths from cardiovascular diseases attributed to stroke and heart disease. Alarming trends underscore the need for a more proactive approach to identifying and mitigating risk factors, including smoking, poor diet, high blood pressure, and sedentary lifestyles. Early diagnosis plays a pivotal role in reducing the damage caused by heart attacks, where clinical methods such as electrocardiography (ECG) and blood tests for cardiac biomarkers like troponin and Creatine Kinase MB (CK-MB) are commonly employed. However, the desire to enhance heart attack detection has given rise to computer-aided systems, particularly machine-learning models that harness diagnostic data and patient information (Aghamohammadi et al., 2019; Reddy et al., 2019). These models are integrated into decision support systems aimed at assisting healthcare professionals in timely and precise diagnoses.

This paper contributes to the ongoing effort to improve heart attack prediction through a comprehensive evaluation of machine learning algorithms (Eladham et al., 2023; Janaraniani et al., 2022; Masethe et al., 2014). By exploring the efficacy of Binary logistic regression (BLR) and Decision Trees, the aim is to identify the most accurate predictive model (Kumar et al., 2022; Maher et al., 2019; Manikandan et al., 2017). The analysis will focus on accuracy, precision, recall, and F-1 scores as key performance metrics (Nayak et al., 2019; Soni et al., 2011).

Additionally, the influence of various features on the predictive outcomes will be investigated using the UC Irvine Machine Learning Repository's heart disease dataset. As the healthcare industry stands on the cusp of a data-driven transformation, the importance of refining predictive models for cardiovascular diseases cannot be overstated (Srinivas et al., 2010). This research contributes to the ongoing quest to harness data analytics and machine learning for early heart attack detection, ultimately improving patient outcomes and reducing the global burden of heart disease.

https://doi.org/10.33436/v34i1y202403

2. Literature survey

The research paper (Dbritto et al., 2016) titled "Improvement of heart attack prediction by feature selection methods" focuses on enhancing the prediction using feature selection algorithms. Among the algorithms that have been explored, support vector machines for classification and Relief (a feature selection method) provided the highest accuracy of 84.81%. The research paper (Obasi et al., 2019) titled "Heart Attack Prediction System" used the Naive Bayes Algorithm to build a classification model with an accuracy of 81.25%.

The papers mentioned herewith reinforce the significance of utilizing machine learning algorithms in the context of heart disease prediction (Takci et al., 2018). The works serve as a valuable reference and validation of the findings and methodologies employed in this study. This study has further worked to build classifiers demonstrating 88% and 92% accuracy.

3. Preliminary data analysis

Based on the analysis of the current data set, the following inferences have been made. The rate of people who suffer from a heart attack is around 54%, comprising 45% of the male population and 75% of the female population. This result suggests that women are more susceptible to heart conditions than men.

Cholesterol levels were analyzed during the study, and several measures of central tendency were noted. For men's cholesterol level, the mean was found to be 240.14 mg/dL and the median 235.0 mg/dL. Women's cholesterol levels were found to be slightly higher, with a mean value of 261.30 mg/dL and a median of 253.0 mg/dL. The analysis concluded that higher cholesterol levels did not correspond with increased susceptibility to heart attack, as the median cholesterol levels of men and women who suffered from heart disease were found to be 235 and 253 mg/dL respectively.

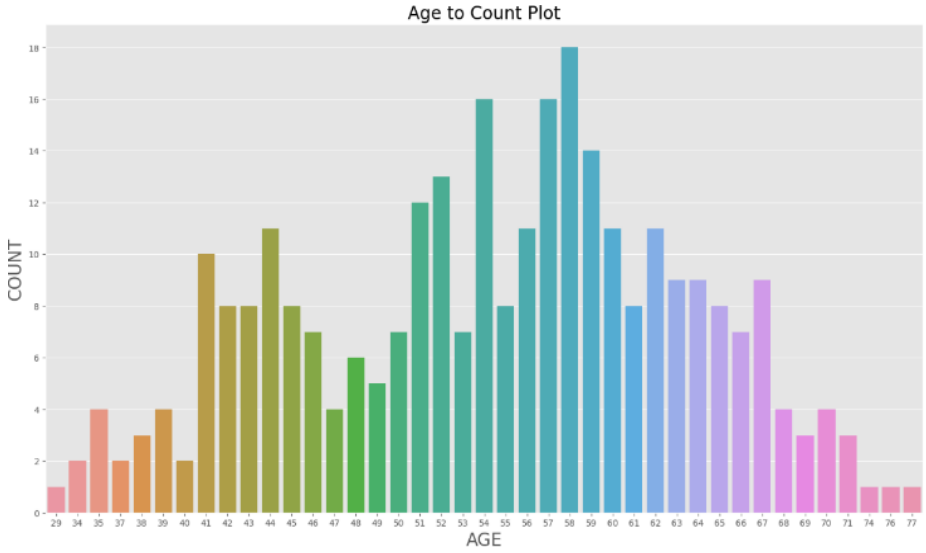

Figure 1. Age to count plot

This paper also examined the effect of age on the risk of cardiovascular disease, as in Figure 1. Based on the given data, it was found that while a higher age did not seem to correlate to an increased risk of heart disease, there were an abnormal amount of heart attacks suffered by those within the 50-65 age group. The median age of male cardiac patients was 51.5, while that of female cardiac patients was around 57. These values are, in fact, lower than the median ages of males and females in the dataset, suggesting that age has an equally inconsequential effect on both sexes.

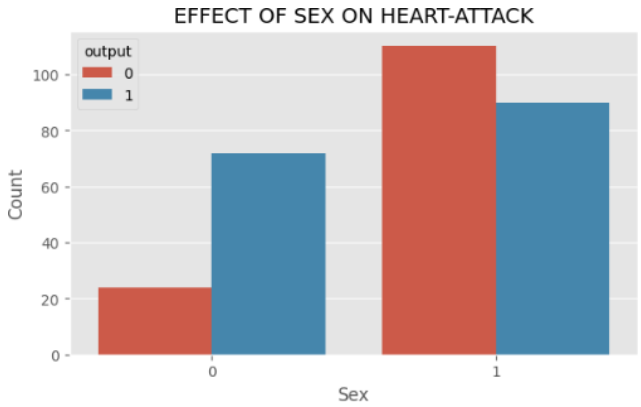

Figure 2. Effect of sex on heart attack

The graph in Figure 2 breaks down the occurrence of heart attack by gender, with 0 representing women and 1 representing men in the X-axis labelled "Sex"; Y-axis, labelled count, shows the number of men and women in the dataset, with a value of "0" representing no heart attack, and 1 representing the patient having suffered a heart attack. This paper's analysis shows that according to our dataset, a significantly higher proportion of women suffer from heart attacks compared to men. However, this is most likely due to skewed data. The total number of women in our dataset is also significantly lower than that of men.Risk of heart attack with relation to age and cholesterol effects:

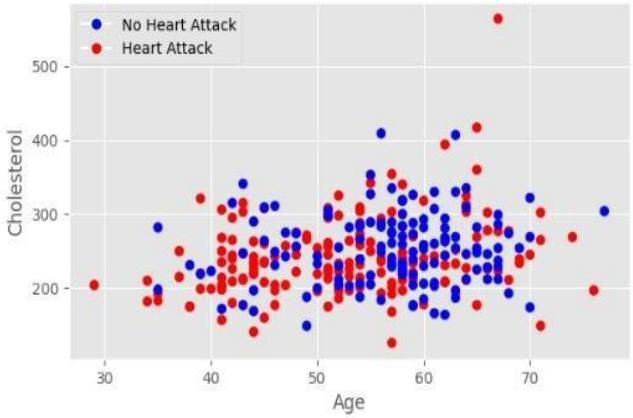

Figure 3. Cholesterol to age scatterplot

The scatter plot of our dataset in Figures 3 and 4 shows no significant relationship between cholesterol levels and a tendency towards heart attack. Based on the analysis of various machine learning approaches, as substantiated in the succeeding sections, two models have been developed using Decision Trees and Logistic Regression.

4. Proposed system

The architecture diagram of the proposed system is given in Figure 4.

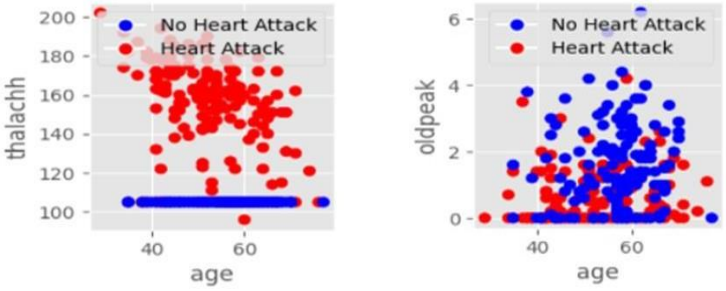

Figure 4. Thalachh to age, oldpeak to age scatterplots

4.1. Examination of various ML approaches based on EDA

4.1.1. KNN approach

The KNN approach can be used for both classification and regression problems. The output of a KNN algorithm is largely dependent on the distances between the neighbouring data points.

As observed from the following scatterplots, it would be difficult to create an accurate model for this dataset using the KNN algorithm due to the following reasons: (i) The number of records in the current dataset is high (296); (ii) For the below identified features, the data points are very closely spaced, and there is very little variation in their distances.

4.1.2. Logistic regression approach

A logistic regression approach is especially suitable for binary classification problems. It is a well-established statistical technique with a long history of use in medical research. Medical datasets usually contain a combination of numerical and categorical values, which can be efficiently handled by logistic regression.

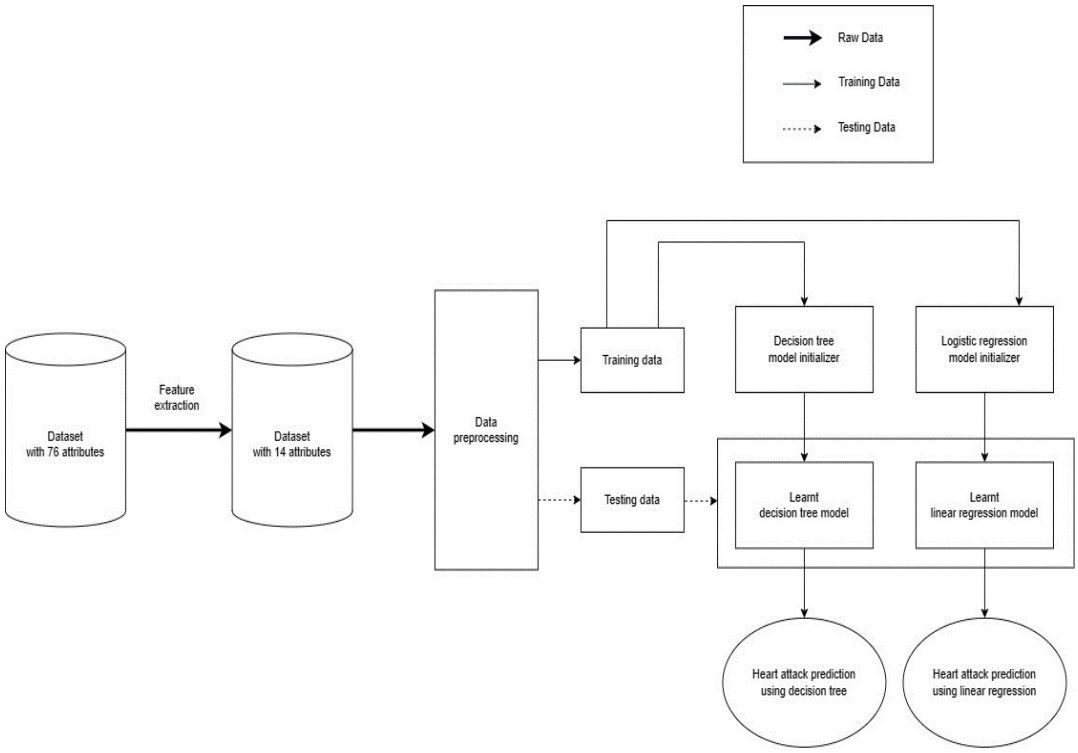

Figure 5. Architecture diagram

4.1.3. Decision tree approach

The decision tree approach is generally used with classification problems. It can also be used with regression ones. A decision tree algorithm is a specifically preferred method to deal with non-linear data while also taking feature-importance measures. Based on our analysis in the previous section, we find that there are two most suitable approaches for model building among the others discussed, as below:

First Approach: Decision Tree - The decision tree algorithm is used to develop a predictive model that can analyze relevant features and accurately classify instances as either prone to or unlikely to experience heart attack.

Second Approach: Logistic Regression - Logistic regression is especially used in dealing with binary classification problems. For our data, the model built with logistic regression displayed 92% accuracy.

5. Dataset

The dataset is taken from the UC Irvine Machine Learning Repository. The features of the dataset are listed in Table 2. The dataset contains 76 attributes and 303 instances, but most of the researchers only make use of a subset of 14 attributes of them. The "goal" attribute in the dataset refers to the presence of heart disease in the patient. It is an integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have focused on simply distinguishing presence (values 1,2,3,4) from absence (value 0). The names and social security numbers of the patients were recently removed from the database and replaced with dummy values.

6. Experiment results and analysis

The results from the two models suggest that the Logistic regression model performs better across all metrics than the Decision Tree model and the KNN model, especially in the case of the model's Precision (97% vs 90%). This suggests that the Logistic regression model is especially useful for diagnosis due to having much fewer false positives compared to the others.

Both models performed relatively well, with the Decision Tree model having an accuracy of 88% and the KNN model having an accuracy of 83% compared to the Logistic regression model's accuracy of 92%, along with F1- Scores of 82%, 89% and 92% respectively.

Table 1. Performance Metrics for Different Approaches

Approach	Accuracy	Precision	Recall	F1-Score
Logistic Regression	92%	97%	88%	92%
Decision Tree	88%	90%	88%	89%
K-Nearest Neighbors	83%	96%	72%	82%

Table 2. Features in the dataset

Sr. no.	Feature	Characteristic representation	Specifics
1	Age	Age	Patients age, in years
2	Sex	Sex	0=female; 1=male
3	Chest pain	Cp	4 types of chest pain (1—typical angina; 2—atypical angina; 3—non-anginal pain; 4—asymptomatic)
4	Rest blood pressure	trestbps	Blood pressure normal (in mm Hg on admission to the hospital)
5	Serum cholesterol	chol	Cholesterol level in mg/dl.
6	Fasting blood sugar	fbs	Fasting blood sugar>120 mg/dl (0—false; 1—true)
7	Rest electrocardiograph	restecg	0—normal; 1—having ST-T wave abnormality; 2—left ventricular hypertrophy
8	MaxHeart rate	Thalach	heart rate achieved which is at its maximum value.
9	Exercise-induced angina	exang	Exercise-induced angina (0—no; 1—yes)
10	ST depression	oldpeak	ST depression induced by exercise relative to rest
11	Slope	slope	slope of the peak exercise ST segment (1—upsloping; 2—fat; 3—down sloping)
12	No. of vessels	ca	No. of major vessels (0–3) coloured by fluoroscopy
13	Thalassemia	thal	3 = normal; 6 = fixed defect; 7 = reversable defect.
14	Num(class attribute)	Class	diagnosis of heart disease status 0—nil risk; 1- high risk;

7. Conclusion and future work

In this study, we have successfully developed and evaluated two predictive models for heart attack prediction: the decision tree model and the logistic regression model. These models have demonstrated promising results, with the decision tree achieving an accuracy of 88 % and the logistic regression model performing even better at 92%. These outcomes underscore the potential of machine learning techniques in assisting healthcare professionals with early heart attack detection.

However, our work is not without its limitations. As the dataset size is small, building deep learning models will not give good results. Further research and refinement are essential to enhance the robustness and applications of these models. In future, the models can be fine-tuned by calculating the energy or the power of the signals and detecting the "distance" between the signal peaks. Exploring additional features or engineering existing ones to capture more nuanced information related to heart health can be done to improve the model's performance.

REFERENCES

Aghamohammadi, M., Madan, M., Hong, J. K. & Watson, I. (2019) Predicting heart attack through explainable artificial intelligence. In *Computational Science – ICCS 2019: 19th International Conference, Faro, Portugal, June 12–14, 2019*. Proceedings, Part II, 19. pp. 633-645. Springer International Publishing.

Dbritto, R., Srinivasaraghavan, A. & Joseph, V. (2016) Comparative analysis of accuracy on heart disease prediction using classification methods. *International Journal of Applied Information Systems*. 11(2), 22-25. doi:10.5120/ijais2016451578.

Eladham, M. W., Nassif, A. B. & AlShabi, M. A. (2023, June) Heart attack prediction using machine learning. In *Smart Biomedical and Physiological Sensor Technology XX*. 12548, 86-93. SPIE. doi:10.1117/12.2664047.

Janaraniani, N., Divya, P., Madhukiruba, E., Santhosh, R., Reshma, R. & Selvapandian, D. (2022, September) Heart Attack Prediction using Machine Learning. In *2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE. pp. 854-860. doi: 10.1109/icirca54612.2022.9985736.

Kumar, A., Rathor, K., Vaddi, S., Patel, D., Vanjarapu, P. & Maddi, M. (2022, August) ECG Based Early Heart Attack Prediction Using Neural Networks. In *2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE. pp. 1080-1083. doi:10.1109/ICESC54411.2022.9885448.

Maher, S., Hannan, S. A., Tharewal, S. & Kale, K. V. (2019) HRV based Human Heart Disease Prediction and Classification using Machine Learning. *International Journal of Computer Applications*. 177(27), 29-34. doi:10.5120/ijca2019919714.

Manikandan, S. (2017, August) Heart attack prediction system. In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, India*. IEEE. pp. 817-820. doi: 10.1109/ICECDS.2017.8389552.

Masethe, H. D. & Masethe, M. A. (2014, October) Prediction of heart disease using classification algorithms. In *Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II WCECS 2014, 22-24 October, 2014, San Francisco, USA*. pp. 25-29.

Nayak, S., Gourisaria, M. K., Pandey, M., & Rautaray, S. S. (2019, May). Prediction of heart disease by mining frequent items and classification techniques. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. IEEE. pp. 607-611. doi:10.1109/ICCS45141.2019.9065805.

Obasi, T. & Shafiq, M. O. (2019, December) Towards comparing and using Machine Learning techniques for detecting and predicting Heart Attack and Diseases. In *2019 IEEE international conference on big data (big data)*. IEEE. pp. 2393-2402. doi:10.1109/BigData47090.2019.9005488.

Reddy, N. S. C., Nee, S. S., Min, L. Z. & Ying, C. X. (2019) Classification and feature selection approaches by machine learning techniques: Heart disease prediction. *International Journal of Innovative Computing*. 9(1). doi: 10.11113/ijic.v9n1.210.

Soni, J., Ansari, U., Sharma, D. & Soni, S. (2011) Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*. 17(8), 43-48. doi:10.5120/2237-2860.

Srinivas, K., Rani, B. K. & Govrdhan, A. (2010) Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering (IJCSE)*. 2(2), 250-255.

Takci, H. (2018). Improvement of heart attack prediction by the feature selection methods. *Turkish Journal of Electrical Engineering and Computer Sciences*. 26(1), 1-10. doi:10.3906/elk-1611-235.

Micheal BERDINANTH is an undergraduate student pursuing a degree in Computer Science and Engineering (CSE) at Sri Sivasubramaniya Nadar College of Engineering.

Samah SYED is an undergraduate student pursuing a degree in Computer Science and Engineering (CSE) at Sri Sivasubramaniya Nadar College of Engineering.

Shudhesh VELUSAMY is an undergraduate student pursuing a degree in Computer Science and Engineering (CSE) at Sri Sivasubramaniya Nadar College of Engineering.

Angel Deborah SUSEELAN is an Assistant Professor in the Department of Computer Science, has 11 years and 6 months of teaching and research experience. She received her Ph.D. from Anna University. She received her M.E in Embedded System Technologies from Anna University of Technology Tirunelveli. She has been awarded first rank in M.E from Anna University of Technology Tirunelveli. She received her B.E. from Jerusalem College of Engineering, Chennai. She is a member of the Machine Learning Research Group of SSN, and she has been consistent in taking part in the International Workshop on "Semantic Evaluation" (SemEval) organized by the Association of Computational Linguistics from 2017. One of the systems developed by her for Fine-Grained Sentiment Analysis on Financial Microblogs and News in SemEval-2017 organized under the umbrella of SIGLEX, the Special interest group on the lexicon of the Association for Computational Linguistics, was ranked third. Also, one of the systems developed by her team was ranked first in ACL 2022. Two of her papers have been awarded the "Best Paper Award" at two different international conferences. She has published 17 papers in journals, 44 papers in international conferences and 3 papers in national conferences. Her research areas include Machine Learning, Natural Language Processing, Data Science, Embedded Systems and Internet of Things.

Rajalakshmi SIVANAIAH is an Assistant Professor in the Department of Computer Science and Engineering and has a total of 19 years of experience in teaching and research. She received her Ph.D. in the area of "Content Boosted Hybrid Filtering for Enhanced Personalization in Recommendation System" from Anna University, Chennai. She completed her M.E degree in Computer Science and Engineering from PSG College of Technology of Bharathiyar University. She did a B.E degree in Computer Science and Engineering from National Engineering College, Manonmaniyam Sundaranar University. She acquired the university's fourth rank in it. She had working experience at Coimbatore Institute of Technology, Coimbatore and MepcoSchlenk Engineering College, Sivakasi, for more than 5 years. She has been working in SSN CE since 2008. Her area of interest includes Data Mining, Machine learning, NLP and Internet Technologies. She has published many papers in various conferences and journals and has attended various workshops and faculty development programs. She has organized various workshops in Natural Language Processing and Embedded Software development. She has organized a Six-day Anna University Sponsored FDTP on "Problem Solving and Python Programming".