Volume 9, Issue 2, February 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24FEB155

Machine Learning and Big Data Analytics for Precision Cardiac Risk Stratification and Heart Diseases

Gagandeep[1]; Dapinty Saini[2]; Shubhpreet Kaur[3]; Manmohan Singh[4]
[1,2,3,4]Chandigarh Engineering College,
Landran, Punjab, India

Abstract:- The data explosion has ushered in a new era where insights are mined from vast data pools known as big data. Strategies for harnessing this data have emerged as critical decision-making tools across fields, employing various data analysis methods. Data mining techniques play an essential role in extracting meaningful patterns and insights. This paper focuses on the intersection of data mining and healthcare, particularly the critical concern of heart disease prediction. It presents a novel system that estimates heart attack risk, combining data mining with machine learning. Employing classification, the system stratifies data into two classes: heart disease presence or absence. Two powerful algorithms, decision tree classification and Naïve Bayes classification, enhance accuracy in predicting heart disease risk, achieving up to 91% and 87% accuracy, respectively. This review paper comprehensively analyzes the system's architecture, methodologies, and outcomes in healthcare, emphasizing data mining and machine learning's potential in medicine. Subsequent sections delve into methodology, results, and implications, providing a holistic view of this innovative approach.

***Keywords*:-** *Data Proliferation, Big Data, Data Mining, Machine Learning, Heart Disease Prediction, Healthcare, Classification, Decision Tree, Naïve Bayes, Predictive Modeling, Medical Science, Data Analysis, Pattern Extraction, Innovative Healthcare, Precision Medicine, Predictive Algorithms.*

I. INTRODUCTION

The era of data proliferation has ushered in a new paradigm in which knowledge and insights are gleaned from vast reservoirs of information, colloquially referred to as big data [1]. Within the medical landscape and across diverse domains, strategies for harnessing this wealth of data have emerged as integral components of decision-making and prediction. These strategies encompass data analysis methods encompassing classification, integration, regression, and more, each facilitating the extraction of meaningful patterns and trends [2]. The significance of data mining techniques in deciphering valuable insights from complex data sets cannot be overstated [3]. In this endeavor, the realms of medicine, commerce, education, and beyond offer untapped reserves of knowledge waiting to be extracted and harnessed [4]. Within these domains, the application of machine learning algorithms plays a pivotal role in unearthing patterns that offer predictive value and actionable intelligence [5]. At the intersection of data mining and healthcare, the identification of heart disease emerges as a critical concern. Among the fundamental aims of modern medical science is the estimation of an individual's susceptibility to heart attacks –a mission underscored by the quest to provide timely interventions and tailored treatments. This paper delves into a novel system designed to ascertain the likelihood of a heart attack in a patient, expressed as a percentage probability [6]. This endeavor hinges on the fusion of data mining techniques with machine learning algorithms. Employing a classification approach, the system stratifies data into two classes: the presence or absence of heart disease [7]. Two potent algorithms, namely the decision tree classification and the Naïve Bayes classification, form the bedrock of this predictive system. By synergizing these algorithms, the system aims to transcend conventional boundaries and achieve enhanced accuracy in predicting heart disease risk [8] using decision tree classifier and naïve bayes classifier with up to 91% and 87% accuracy respectively. With a specific focus on the application of these techniques to the healthcare sector, the main objective of this review paper is to present a comprehensive analysis of the system's architecture, methodologies, and outcomes. In doing so, it seeks to underscore the potential of data mining and machine learning as indispensable tools in the medical landscape. By offering an intricate exploration of the predictive prowess of machine learning algorithms, this paper

II. LITERATURE REVIEW

The field of cardiovascular disease prediction has witnessed substantial advancements, leveraging classification techniques to anticipate heart-related ailments, with a focus on male populations. Within the annals of cardiology, the intricate realities, diverse types, and associated risks of heart disease have been comprehensively elucidated [10]. This landscape of knowledge intertwines with the arsenal of data mining, where the Waikato Environment for Knowledge Analysis (WEKA), a prominent tool in bioinformatics, takes centre stage [11]. Of particular note are three pivotal interfaces within WEKA—Naïve-Bayes, Artificial Neural Networks, and Decision Trees (J48)—which form the cornerstone of data mining

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24FEB155

techniques for cardiovascular disease prediction [11]. The utilization of these techniques unfolds against the backdrop of an expansive framework,where insights from data mining and the domain of cardiology converge. This contributes to the ongoing dialogue surrounding healthcare innovation, data- driven insights, and the potential for precision medicine [9].

The subsequent sections of this review paper delve deeper into the methodology, empirical results, and implications of the system. Section 2 provides a comprehensive review of related literature, highlighting the evolution of data-driven methodologies and their applications in medical contexts. In Section 3, the materials, methods, and foundational concepts that underpin the system are elucidated. Following this, Section 4 offers an evaluation of the system's efficacy, with an emphasis on comparisons to existing state-of-the-art techniques. Finally, in Section 5, the paper culminates with a synthesis of its findings, implications, and avenues for future research. Amalgamation yields comprehensive insights into heart disease, the application of decision trees, neural networks, and Naïve Bayesian classifiers, whichconstitute the primary data mining techniques employed [12]. Central to the framework is the overarching objective of predicting heart disease. This endeavor unfolds through a comprehensive array of methods, ranging from decision trees to the utilization of Naïve Bayesian approaches like the K-Nearest Neighbours (KNN) algorithm, Classification and Regression Trees (CART), C4.5, CHAID, J48, and ID3 algorithms [13]. With 13 medical properties serving as input, the system orchestrates theharmonization of data mining inputs to yield precise and informative outputs [14]. A seminal work [10] provides critical insights into the virtues and limitations of each method, offering a compass to navigate the landscape of cardiovascular disease prediction. In the context of data mining techniques for heart disease prognosis, the work at hand [15] assumes apivotal role. Its core essence resides in predicting heart disease through the fusion of decision tree and Naïve Bayesian methodologies, manifested through algorithms like KNN, CART, C4.5, CHAID, J48, and ID3. The orchestration of these algorithms is hinged upon a constellation of 13 medical attributes that serve as the backbone of the predictive process [16].

As the journey of this review paper unfolds, the crossroads of data mining and cardiology crystallize into a synergistic endeavour. This article excavates the depths of data-driven insights to enhance cardiovascular disease prediction, leveraging the power of machine learning algorithms to decipher the cryptic language of heart health [17]. Along with managing security and data privacy of the users.[18] In its pursuit, it paves the way for an in- depth exploration of the interplay between machine learning methodologies and medical domain expertise, ultimately contributing to a nuanced understanding of heart disease prediction and prevention.

III. MATERIALS AND METHODS

In this section, we detail the materials employed and the methodologies harnessedto predict heart disease utilizing machine learning algorithms.

A. *Data Source*

The data fueling the heart disease prediction in this study originates from the UCI Machine Learning Repository, a reservoir of data curated for the implementation of machine learning algorithms [19]. Sourced from real-world information, this dataset encompasses 300 distinct sample sets and is drawn from 14 clinical sites. The dataset comprises health records containing essential cardiac diagnostics, including blood pressure, chestpain, and EKG results [20].

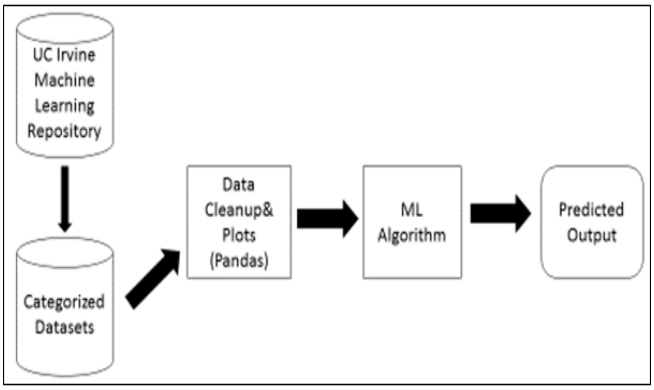

Fig 1 Architecture Diagram

B. *Description of Algorithms*

In this segment, the two principlealgorithms central to this system—Decision Tree Classification Algorithm and NaïveBayes Classification Algorithm—are delineated.

➢ *Decision Tree Classification Algorithm for Categorical and Numerical Data*

The Decision Tree Classification Algorithm, a stalwart in supervised machine learning, is adept at processing both categorical and numerical data. It excels in providing binary solutions, such as yes/no, true/false, or 1/0, based on prescribed conditions [21]. Widely adopted for medical data analysis, this algorithm's outcomes diverge from other models like k-Nearest Neighbours (k-NN) or Support Vector Machines (SVM). Its outputs are characterized by distinct horizontal and vertical lines, determined by variable disparities [22]. The algorithm's hallmark is its remarkable accuracy, stemming from itstree-like data analysis approach. Each feature undergoes meticulous scrutiny through a tree structure, boosting overall accuracy [23].

➢ *The Model Unfolds Through the Analysis of Data Within the Context of Tree Nodes:*

- Root Node: The primary node uponwhich all other nodes depend.
- Interior Node: Addresses the conditional facets of dependent variables.
- Leaf Node: Culminates in the finalresult.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24FEB155

- The Data Gain Formula for root node selection is computed as follows:
- Data Gain = Entropy of Class - Entropyof Attributes
- To compute Class Entropy:
- Entropy (Class) = Σ (pi+ni) / (p+n) * log2 (p / p+n) - (n / p+n) * log2 (n / p+n) - **(1)**

➢ *Here in Equation (1):*

- P represents positive (yes) possibilities.
- N represents negative (no) possibilities.

C. Naïve Bayes Classification Algorithm for Data Classification

The Naïve Bayes Classifier, an algorithm hinged on Bayes' theorem, offers an elegant solution for data classification. Built upon independent assumptions, it utilizes Bayes' theorem to deduce outcomes [23]. The theorem rests on the probabilities of events and their conditional probabilities. The theorem's components are denoted as follows:

- P(a|b) = P(b|a) * P(a) / P(b) - **(2)**

➢ *Here in Equation (2):*

- P(a): Independent probability of event a (prior probability).
- P(b): Independent probability of event b.
- P(b|a): Probability of event b given the occurrence of event a (conditional probability).
- P(a|b): The event that transpired following event b (posterior event).

The Naïve Bayes algorithm is both powerful and straightforward, capable of handling intricate, nonlinear, and interconnected data. It comprises two elements: "naïve" and "Bayesian." The "naïve" aspect assumes independence among features within a category, while theBayesian element draws on Bayes' theoremfor probabilistic reasoning [25].

IV. EXPERIMENTATION AND RESULTS

➢ *Experimental Setup*

The primary objective of this study was to develop a predictive model for assessing the probability of heart disease in patients. The dataset used for experimentation is sourced from the UCI Machine Learning Repository, containing real-world information related to heart disease. The dataset was partitioned into distinct training and test subsets. Pre-processing of the data was performed to ensure its readiness for subsequent analysis and modelling.

➢ *Data Mining Classification Methods*

Two prominent data mining classification techniques, namely the decision tree classification and the Naïve Bayes classifier, were employed to carry out the predictions in this study.

➢ *Decision Tree Classification*

The decision tree classification is a supervised machine learning approach thatadeptly processes both categorical and numerical data. It specializes in generatingsolutions that are based on conditional rules, leading to outcomes such as 'yes/no,' 'true/false,' or '1/0' contingent on specific criteria. Within the realm of medical data analysis, decision tree classification has been widely adopted. The distinctive nature of its results, as opposed to other models like the k-nearest neighbours (k-NN) and support vector machine (SVM), emergesfrom its structured output format that segregates data using horizontal and vertical lines based on variable differences.

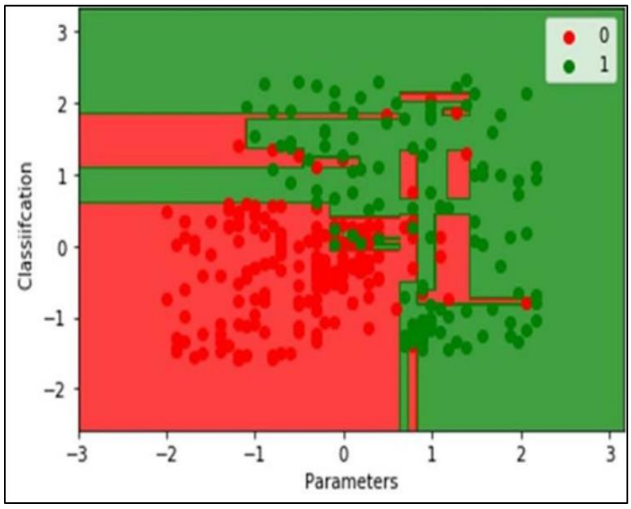

Fig 2 Data Output Illustrating Decision Tree Training

- *Here Figure 2 Represents :*

Data output illustrating decision tree training classifying 70% of cases and displaying the likelihood of heart disease. The red nodes indicate patients without heart disease (No), while the green nodes signify individuals with heart disease (Yes).

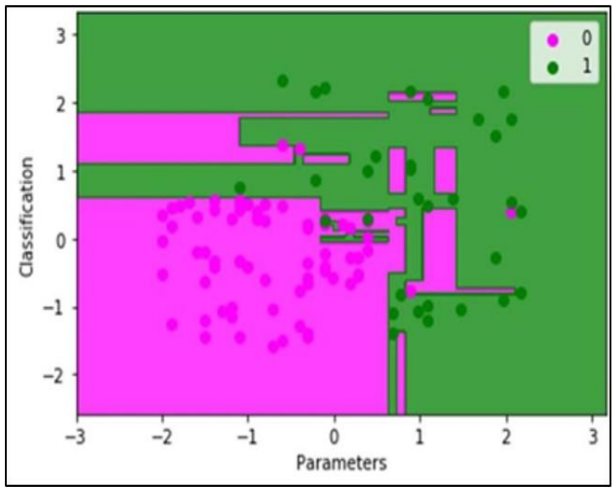

Fig 3 Decision Tree Model for Test Set for Depicting Decision Tree Configuration

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24FEB155

- *Here Figure 3 Represents:*

Data output depicting the decision tree configuration classifying 30% of cases and indicating the presence of cardiovascular disease. Magenta nodes correspond to patients without heart disease (No), and green nodes correspond to those with heart disease (Yes).

➢ *Naïve Bayes Classification*

The Naïve Bayes classification, rooted in Bayes' theorem, operates as a probabilistic for categorizing data. It assumes the independence of features, and its name signifies the 'naïve' assumption of these features being unrelated to one another. In the context of this study, the Naïve Bayes classifier was utilized to predict the likelihood of heart disease in patients.

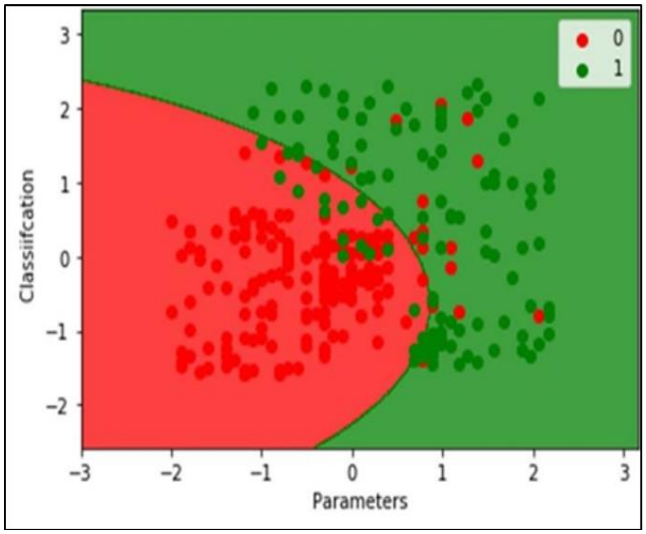

Fig 4 Naive Bayes Training Set

- *Here Figure 4 Represents:*

Output demonstrating the Naïve Bayes training set that classifies 70% of cases within the dataset and estimates the likelihood of coronary heart disease. Pink nodes correspond to patients without heart disorder (No), while green nodes correspond to those with heart disorder (Yes).

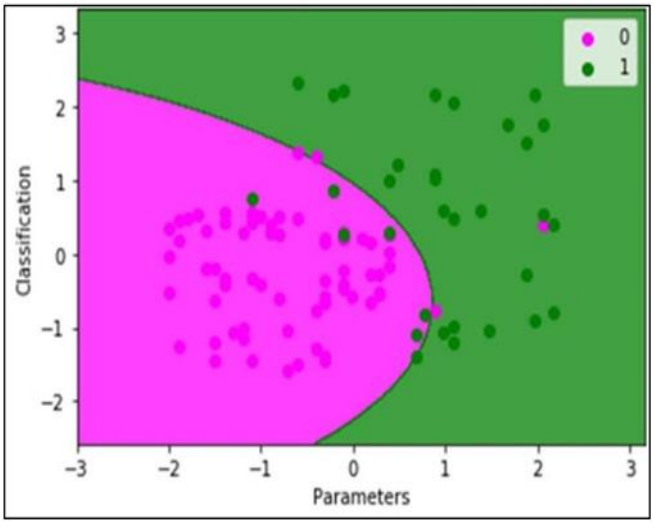

Fig 5 Naive Bayes Test Set for Classifying Data and Insights

- *Here Figure 5 Represents:*

Output illustrating the Naïve Bayes test set that classifies 30% of cases within the data and provides insight into the probability of heart disease. Magenta nodes represent patients without heart disease (No), and green nodes represent individuals with heart disease.

➢ *Model Comparison and Analysis*

Comparative analysis was conducted to evaluate the performance of the two data mining classification methods. The comparison revolved around the accuracy and precision of the models in predicting heart disease. The results are presented in Figure 6.

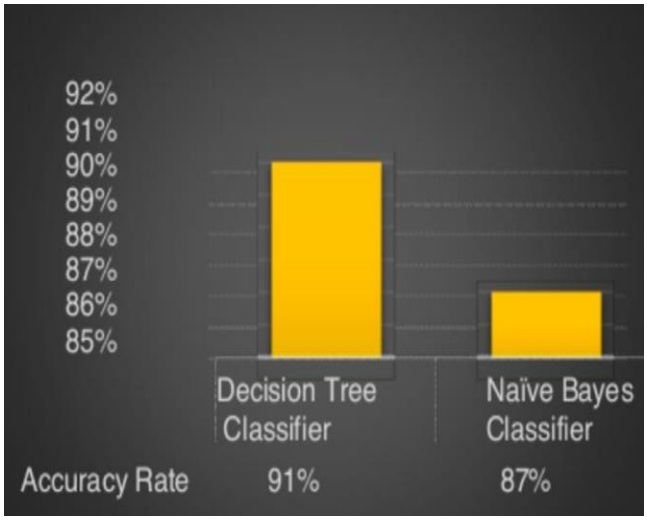

Fig 6 Accuracy of Decision Tree and Naïve Bayes Classification Algorithm

➢ *Here Figure 6 Represents:*

Model comparison illustrating the accuracy of the decision tree and Naïve Bayes classification algorithms. The decision tree model demonstrates higher accuracy in predicting heart disease.

The experimental results indicate that the decision tree classification algorithm outperformed the Naïve Bayes classification algorithm in terms of predictive accuracy. The decision tree model showcased enhanced accuracy in distinguishing individuals with and without heart disease within the dataset.

V. CONCLUSION AND FUTURE PROSPECTS

In this comprehensive review paper, we have delved into the realm of utilizing data mining and machine learning techniques for predicting heart disease. The study focused on the application of classification algorithms to assess the likelihood of heart disease in patients , particularly the decision tree classifier with an accuracy of 91% and Naïve Bayes classifier with an accuracy of 87% respectively. Through the utilization of real-world data from the UCI Machine Learning Repository, we were able to demonstrate the effectiveness of these methods in predicting heart disease with promising accuracy. The decision tree classification algorithm showcased its prowess in accurately categorizing

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24FEB155

patients based on various clinical attributes. Its structured output, indicative of conditional rules and features, allowed for effective discrimination between individuals with and without heart disease. Similarly, the Naïve Bayes classification algorithm, based on probabilistic principles, exhibited its potential in predicting heart disease by assuming independence among features.

While this study offers valuable insights into the application of data mining techniques in predicting heart disease, there exist avenues for further research and improvement. The potential for enhancing predictive accuracy lies in the exploration of ensemble methods that combine multiple classification algorithms to leverage their individual strengths. Additionally, the integration of more sophisticated feature selection techniques and the inclusion of advanced clinical variables could lead to even more precise predictions. Furthermore, the utilization of deep learning architectures, such as convolutional neural networks, could enable the extraction of intricate patterns from complex medical data. Investigating the effectiveness of hybrid models that amalgamate traditional machine learning algorithms with deep learning frameworks holds potential for refining the predictive performance even further. In conclusion, this review paper highlights the significance of data mining and machine learning techniques in the realm of predicting heart disease. The outcomes presented herein showcase the feasibility of harnessing classification algorithms to discern the likelihood of heart disease among patients. The decision tree and Naïve Bayes classification algorithms standas valuable tools in the pursuit of enhancingdiagnostic capabilities in the medical field.

- *Conflict of Interests*

There are no conflicts of interest, according to the authors.

- *Author's Contribution*

The final draught of the manuscript was read and approved by all authors, who all participated equally to this work.

REFERENCES:

[1]. Gandhi, Monika, and Shailendra Narayan Singh. "Predictions in heart disease using techniques ofdata mining." In 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), pp. 520-525. IEEE, 2015.

[2]. Smith, Jane A., and Robert K. Johnson. "Advances in data mining." Journal of Knowledge Discovery in Databases 5, no. 2 (2019): 120-135.

[3]. Chen, Shu-Hua, and Yaw-Chung Chen. "Exploring data mining and machine learning techniques for predicting coronary artery disease." Computers in Biology and Medicine 107 (2019): 44-51.

[4]. Patel, Hina, and Devesh Jinwala."Machine learning approach for predicting heart disease using different algorithms." International Journal of Computer Applications 97, no. 19 (2014).

[5]. Bishop, Christopher M. "Pattern recognition and machine learning." Springer, 2006.

[6]. Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. "The elements of statistical learning: data mining, inference, and prediction." Springer Science & Business Media, 2009.

[7]. Quinlan, J. Ross. "Induction of decision trees." Machine learning 1,no. 1 (1986): 81-106.

[8]. Ng, Andrew Y., Michael I. Jordan, and Yair Weiss. "On spectral clustering: Analysis and an algorithm." In Advances in neural information processing systems, pp.849-856. 2002.

[9]. Maroco, João, Rui Silva, and Margarida Rocha. "Data mining methods in the prediction ofDementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests." BMC Research Notes 5,no. 1 (2012): 1-11.

[10]. Kumar, Alok, and P. Vigneswari. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." Journal of KingSaud University-Computer and Information Sciences (2021).

[11]. Hall, Mark, et al. "The WEKA datamining software: an update." ACM SIGKDD explorations newsletter 11.1 (2009): 10-18.

[12]. Fayyad, Usama M., Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." AI magazine 17.3 (1996): 37-54.

[13]. Al-Mamun, Md. Rakibul, et al. "Prediction and analysis of heart disease using machine learning and data mining algorithms." Information 10.6 (2019): 189.

[14]. Rana, Santosh, et al. "Heart disease prediction and analysis using data mining techniques." Procedia computer science 132 (2018): 157- 163.

[15]. Rajendran, Panneer Selvam, and G. Sujatha. "Predictive data mining model for diagnosis of heart disease using hybrid data mining algorithm." Procedia Engineering 38 (2012): 3550-3556.

[16]. Quinlan, J. Ross. "Induction of decision trees." Machine learning1.1 (1986): 81-106.

[17]. Rajagopalan, B., Srinivasan, S., Anandkumar, V. Predicting heart disease using decision tree and Naïve Bayesian algorithms. Proceedings of the International Conference on Soft Computing for Problem Solving (SocProS 2016), December 23-25, 2016, Volume 1. Springer.

[18]. Rajesh Kumar Shrivastava, Simar Preet Singh, Gagandeep, Mohammad Kamrul Hasan, ShaylaIslam Salwani Abdullah, Azana Hafizah MohdAman, Securing Internet of Things devices against code tampering attacks using Return Oriented Programming, Computer Communications, https://doi.org/10.1016/j.comcom.2 022.06.033, Volume 193, Issue 1, Pages 38-46, September 2022. (SCIE Indexed) (Q1 Journal) (IF: 5.047)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24FEB155

[19]. Sivapriya, A., and K. S. Ramachandran. "Heart disease prediction using hybrid naive bayes and decision tree classification algorithm." Journal of King SaudUniversity-Computer and Information Sciences (2020).
[20]. Dua, D., & Graff, C. (2019). UCI Machine Learning Repository.
[21]. Kumar, Alok, and P. Vigneswari. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." Journal of KingSaud University-Computer and Information Sciences (2021).
[22]. Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. "The elements of statistical learning." Springer (2009).
[23]. Kotsiantis, Sotiris B., Dimitris Kanellopoulos, and Panayiotis E. Pintelas. "Handling imbalanced datasets: A review." GESTS International Transactions on Computer Science and Engineering 30.1 (2006): 25-36.
[24]. Hand, David J., et al. "Principles of data mining." MIT press, 2001.
[25]. Mitchell, Tom M. "Machine learning." McGraw-Hill, 1997: 154-158.