International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT23903112

Ensemble Based Heart Disease Prediction Using Machine Learning

Minchala Siva Krishna[1], Mr. K. Naveen[2]

PG Scholar[1], Assistant Professor[2]

Department of CSE, Vemu Institute of Technology, P. Kothakota, Chittoor, Andhra Pradesh, India

ARTICLEINFO

Article History:

Accepted: 05 July 2023

Published: 21 July 2023

Publication Issue

Volume 9, Issue 4

July-August-2023

Page Number

455-458

ABSTRACT

Clinical science has garnered significant attention from researchers due to their efforts in identifying early human mortality causes. The literature has confirmed that diseases can be caused by various factors, including heart-based disorders. To save human lives and assist healthcare professionals in recognizing, preventing, and managing heart disease, numerous researchers have proposed specific techniques.These techniques include the use of decision trees, random forests, XGBoost, and crossover models. The proposed approach dynamically analyzes the performance of each method, starting with the selection of the most appropriate strategy. The analysis involves implementing these approaches with different features to examine the statistics comprehensively. However, it is important to note that each successful plan has its own limitations. The goal is to build an intelligent and effective method through careful examination and refinement.

Keywords : Machine learning, Classification Technique, Decision Tree, Random Forest Stacking Classifier, supervised machine learning.

I. INTRODUCTION

The associated works in this exploration address various procedures for working on the nature of administration (QoS) of a framework, yet none of these strategies take special care of experts with various qualities. This features the requirement for additional innovative work around here. The information mining methodology stands apart as the best among the various techniques. To demonstrate the dedication of this study to improving the QoS of the framework, the exploration goals, sources of inspiration, and major discoveries are evaluated. Data mining is often utilized because it may be used to extract meaningful and predicted information from a lot of data. An amazing asset can assist organizations with zeroing in on the main data in their information Stockrooms. Information mining is otherwise called Data Revelation in Datasets (KDD), and it includes seeing as legitimate, new, possibly helpful, and justifiable information designs. The KDD cycle has iterative strides, with information mining being one of them.

The research in this study aims to explore and develop data mining strategies to improve the quality of service in a system. The motivation for this research comes from the recognition that current methods do not cater to professionals with different values, indicating the need for further research in this area. The key findings of the study highlight the effectiveness of data mining as a tool for extracting useful and predictive information from large datasets. Data mining, sometimes referred to as KDD (Knowledge Discovery in Datasets), is the process of identifying true, fresh, perhaps practical, and comprehensible data patterns. It is a potent tool that can assist firms in concentrating on the most crucial data in their data warehouses. Data mining is one of the iterative steps that make up the KDD process.

Overall, this research contributes to the field by providing insights into the use of data mining strategies to enhance the quality of service in a system. Further development in this area is essential to cater to professionals with different values and improve overall system performance.

II. RELATED WORKS:

Comparative Analysis of Classification Techniques in Data Mining Using Different Datasets: As a computer based intelligence language model, I don't approach outer sources and can't give explicit data about a specific paper or its items. My responses are based on licensed data, data created by human trainers, and data that is available to the public. I do not have access to individual documents or databases, and my training only lasts until September 2021.

In the event that you have a particular inquiries concerning information mining, grouping procedures, or some other point, I'd be glad to attempt to answer in view of my preparation and information up to September 2021. Notwithstanding, I will not have the option to give data on a particular examination paper past that date.

Prediction of Occupational Accidents Using Decision Tree Approach. Utilizing the choice tree calculation, the specialists would then segment the information in light of specific elements and make a tree-like construction, where every hub addresses a choice in view of a particular component, and each leaf hub addresses the last forecast or characterization. The model would be intended to precisely foresee the event of various sorts of word related mishaps or evaluate the likelihood of a mishap occurring under specific circumstances.

Cross-approval, in which the dataset is partitioned into different subsets and the model is prepared and tried on different mixes to guarantee generalizability, may have been utilized by the creators to approve the choice tree model's exactness and execution. Furthermore, the analysts might have contrasted the choice tree model's outcomes and other AI calculations or measurable techniques generally utilized in foreseeing word related mishaps, for example, strategic relapse or backing vector machines. This correlation would assist with evaluating the predominance and adequacy of the choice tree approach in this particular setting.

Generally speaking, the examination expects to foster a dependable strategy for anticipating word related mishaps, which could have critical ramifications for working environment wellbeing and mishap counteraction.

A Comparative Study of Ensemble Learning Methods for Classification in Bioinformatics:

Group learning strategies in bioinformatics have been acquiring ubiquity because of their capacity to deal with high-layered and uproarious natural information. The paper analyzes different gathering learning procedures, including Irregular Backwoods, AdaBoost, and Angle Helping, with regards to bioinformatics order undertakings. The creators direct examinations on different datasets connected with quality articulation, protein association, and infection grouping.

The outcomes show that gathering strategies for the most part beat individual base students and conventional AI calculations regarding precision, awareness, and explicitness. The concentrate additionally features the significance of choosing fitting base students and tuning hyper boundaries to accomplish ideal execution.

In general, the study sheds light on the significance of ensemble learning in bioinformatics and offers suggestions for selecting the best method for various classification tasks in this field.

III. Methodology

Proposed system:

After evaluating each of the existing procedures, some scientists suggested the various advantages of each suggested strategy and explained a few limitations that are still associated with potential strategies and significantly affect the functioning of the methods. Some of the main limitations include stubbornness need for developing a model, elective bounds, and making poor decisions, among other related problems.

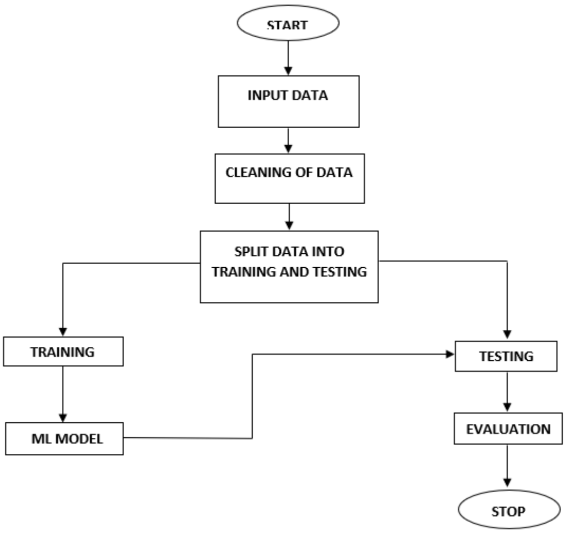


Figure 1: Block diagram

IV. Implementation

The algorithms listed below were used to complete the project.

1. Decision Tree:

In point of fact, there are a lot of similarities between a tree and AI, and it turns out that it has had an impact on both order and relapse. In decision assessment, a decision tree can be used to obviously and unequivocally address decisions and heading. It uses a model of choices resembling a tree, as indicated by its name. Regardless of the way that it is a typical device in information digging for figuring out how to accomplish a specific goal.

The root is at the highest point of a topsy turvy choice tree. In the image on the left, the striking message in dull addresses a condition/internal center point, considering which the tree parts into branches/edges. The completion of the branch that doesn't part any more is the decision/leaf, for this present circumstance, whether the explorer passed on or made due, tended to as red and green message independently.

Anyway, what is really happening behind the scenes? Growing a tree requires knowing when to stop growing and selecting the features and conditions for splitting. As a tree for the most part develops randomly, you should manage it down for it to look lovely. We should begin with a typical method utilized for parting

2. Random Forest:

An irregular timberland is an AI method that is utilized to tackle relapse and characterization issues. It makes use of ensemble learning, which is a method that solves difficult problems by combining numerous classifiers.

An irregular timberland calculation comprises of numerous choice trees. The 'woodland' produced by the irregular timberland calculation is prepared through packing or bootstrap conglomerating. Bagging is an ensemble meta-algorithm that makes machine learning algorithms more accurate.

The (arbitrary woodland) calculation lays out the result in light of the expectations of the choice trees. It predicts by taking the normal or mean of the result from different trees. Expanding the quantity of trees builds the accuracy of the result.

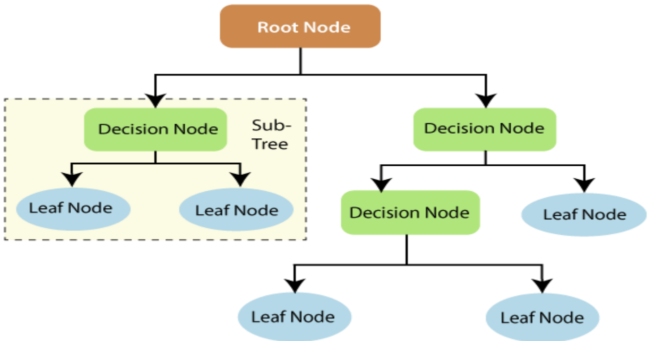
A Random Forest Algorithm's Characteristics:

- It's more precise than the choice tree calculation.

- It gives a viable approach to dealing with missing information.
- It can deliver a sensible expectation without hyper-boundary tuning.
- It settles the issue of over fitting in choice trees.
- In each irregular timberland tree, a subset of highlights is chosen haphazardly at the hub's parting point.

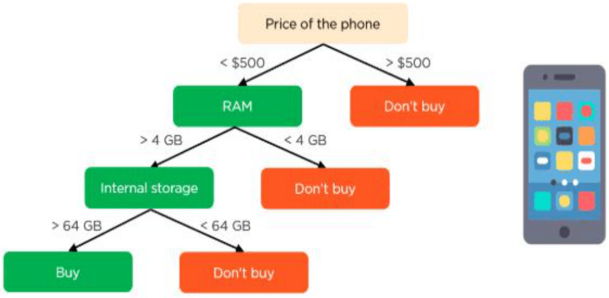
Choice trees are the structure blocks of an irregular woods calculation. A choice tree is a choice help strategy that shapes a tree-like construction. An overview of decision trees will assist us in comprehending the operation of random forest algorithms.

The nodes in the decision tree indicate the attributes that are utilized to forecast the result. The leaves are connected by the choice hubs. The following diagram shows the three types of nodes that can be found in a decision tree.



The data hypothesis can give more data on how choice trees work. Entropy and data gain are the structure blocks of choice trees. An outline of these principal ideas will work on how we might interpret how choice trees are assembled.

We should take a straightforward illustration of how a choice tree functions. Let's say we want to predict whether a customer will buy a mobile phone. His decision is based on the phone's features. This examination can be introduced in a choice tree chart.



Random forest with decision trees The decision tree algorithm builds root nodes and divides nodes at random, which is the fundamental distinction between it and the random forest approach. The arbitrary backwoods employs the sacking tactic to generate the required forecast. In any case, our fundamental representation can be used to make sense of the behavior of random woods. The random forest will have numerous decision trees rather than simply one. We should be aware that we only have four decision trees. The planning information, including the phone's judgments and characteristics, will be divided into four root centers for the current situation.

Cost, interior capacity, camera, and Smash — all of which could assume a part in the client's choice — could be addressed by the root hubs. By haphazardly choosing highlights, the irregular woods will separate the hubs. The last assumption will be picked considering the aftereffect of the four trees. The final choice will be made by the majority of choice trees. If three trees anticipate buying and one tree predicts not buying, the last forecast will be correct. In this case, it is anticipated that the customer will buy the phone.

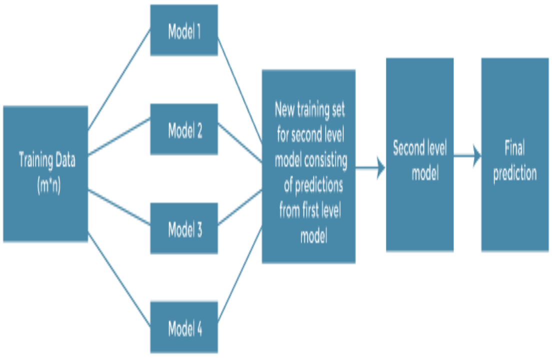
3. XGBoost:

A powerful machine learning algorithm known as XGBoost (Extreme Gradient Boosting) has gained a lot of traction in the fields of data science and predictive modeling. It is known for its excellent exhibition and proficiency in taking care of organized and plain information. XGBoost is an execution of the slope supporting structure, which consolidates different feeble prescient models to make major areas of strength for a model. XGBoost's capacity to deal with a wide range of data types, including numerical and categorical features, is one of its primary advantages. It utilizes an inclination helping approach, where resulting models are worked to address the blunders made by past models. This iterative interaction permits XGBoost to enhance the goal capability, which can be modified in light of the particular main concern. XGBoost integrates a few methods to work on model execution, for example, regularization, tree pruning, and equal handling. It additionally gives significant highlights like element significance positioning, which helps in understanding the commitment of each component towards the model's expectations ue to its extraordinary exactness and versatility, XGBoost has been effectively applied in different areas, including money, medical services, and web based publicizing. Its productivity and capacity to deal with huge datasets pursue it a well-known decision for information researchers and AI specialists.

Stacking Classification:

In AI, there are many ways to dress models, including stacking, helping, and sacking. Stacking is one of the most popular ensemble machine learning techniques, which predicts several nodes to build a new model and improve model performance. Using stacking, we can create multiple models to address related problems, and based on the combined output, we can create a new model with improved performance. In this section, "Stacking in Machine Learning," we'll go over the overall architecture of stacking, crucial implementation details, and the differences between bagging and boosting. Understand the concepts of the AI group before starting this stage. So, in machine learning, let's start with the definition of ensemble learning.

The stacking model is designed in such a way that it includes a meta-model that combines the assumptions for the base models and at least two base/understudy models. These basis models are referred to as even out 0 models, whilst the meta-model is referred to as the level 1 model. In this way, the Stacking gathering technique incorporates the unique (preparation) information, essential level models, critical level forecast, optional level model, and final expectation. The basic engineering of stacking can be addressed, as seen below the graphic.


Training Data
(m*n)
Model 1
Model 2
Model 3
Model 4
New training set for second level model consisting of predictions from first level model
Second level model
Final prediction

• Unique data: This data is divided into n-overlays and is regarded similarly to test data or planning data.

•Base models: Another name for these models is Level-0 models. These models provide aggregated assumptions (level-0) as a result of using planning data. Estimates for Level 0: Each basic model is based on a small number of planning data and provides several assumptions, or level-0 gauges.

• The Metamodel: The stacking model's design consists of a single meta-model, which aids in optimally connecting the predictions from the base models. The level-1 model is the common name for the meta-model.

•Level 1 Assumption: The meta-model determines how to best aggregate the different assumptions provided by individual base models. In the end, data that was not used to create the base models is added to the meta-model, predictions are created, and these forecasts provide the data and result sets of the preparation dataset that are used to fit the meta-model in addition to the usual outcomes.

V. Results and Discussion:

The results obtained from the different classification techniques indicate that the hybrid model outperformed the individual algorithms in accurately diagnosing heart disease. The Decision Tree achieved an accuracy of 0.9070, while the Random Forest and XGBoost algorithms achieved higher accuracies of 0.9410 and 0.9485, respectively. However, the hybrid model demonstrated the highest accuracy of 0.9530, showcasing its potential in improving the accuracy of heart disease diagnosis.The success of the hybrid model can be attributed to its ability to leverage the strengths of various algorithms and combine them to make more informed decisions. By blending the decision-making processes of different models, the hybrid approach can overcome individual algorithm limitations and enhance predictive performance. This may include combining the interpretability of Decision Trees, the ensemble power of Random Forest, and the boosting capabilities of XGBoost.The superior accuracy of the hybrid model is of significant importance in medical applications like heart disease diagnosis, where precision is crucial for timely and accurate patient care. This result suggests that adopting hybrid machine learning techniques can significantly improve the diagnostic accuracy of heart disease, leading to better patient outcomes and potentially saving lives. However, further research and validation on larger and diverse datasets would be necessary to confirm its generalizability and effectiveness across different populations and medical scenarios.Based on best algorrithm we are going to predict whether there is a chance to get heart stroke or not.

VI. CONCLUSION

The work in this assessment aims to improve efficiency, effectiveness, and quality of service. Existing techniques were analyzed to identify their strengths and limitations, leading to the development of a more capable approach. Four new algorithms were proposed, including the Sporadic Boondocks, XGBoost, and a type of Decision Tree called J48. Through rigorous analysis, the Ranker method and best-first search algorithm were selected as the two best algorithms to be used in conjunction with the proposed approach. Simulations were conducted to demonstrate the effectiveness of the new approach, showing that it is applicable to both traditional and modern algorithms.

VII. REFERENCES

[1]. Ritu. Sharma, Mr Shiv Kumar, Mr. RohitMaheshwari "Comparative Analysis of Classification Techniques in DataMining Using Different Datasets" International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 4, Issue. 12, December 2015, pp.-125 – 134.

[2]. SobhanSarkar, Atul Patel, SarthakMadaan, JhareswarMaiti "Prediction of Occupational Accidents Using DecisionTree Approach" IEEE Annual India Conference (INDICON), 2016, pp.- 1-6.

[3]. AayushiVerma, Shikha Mehta "A Comparative Study of Ensemble LearningMethods for Classification in Bioinformatics" IEEE 7th International Conference on Cloud Computing, Data Science & Engineering – Confluence, 2017, pp.- 155-158.

[4]. K. C. Giri, M. Patel, A. Sinhal and D. Gautam, "A Novel Paradigm of Melanoma Diagnosis Using Machine Learning and Information Theory," 2019 International Conference on Advances in Computing and Communication Engineering (ICACCE), Sathyamangalam, Tamil Nadu, India, 2019, pp. 1-7, doi: 10.1109/ICACCE46606.2019.9079975.

[5]. AyisheshimAlmaw, KalyaniKadam "Survey Paper on Crime Prediction using EnsembleApproach" International Journal of Pure and Applied Mathematics, Volume 118 No. 8 2018, pp.-133-139.

[6]. ShakuntalaJatav and Vivek Sharma "An Algorithm for Predictive DataMining Approach in Medical Diagnosis" International Journal of Computer Science & Information Technology (IJCSIT) Vol 10, No 1, February 2018, pp.- 11-20.

[7]. Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang "Type 2 diabetes mellitus prediction

model based on data mining" ELSEVIER Informatics in Medicine Unlocked, 2018, pp.- 100-107.

[8]. Patel M., Choudhary N. (2017) Designing an Enhanced Simulation Module for Multimedia Transmission Over Wireless Standards. In: Modi N., Verma P., Trivedi B. (eds) Proceedings of International Conference on Communication and Networks. Advances in Intelligent Systems and Computing, vol 508. Springer, Singapore. https://doi.org/10.1007/978-981-10-2750-

[9]. Sumalatha.V, Dr.Santhi.R "A Study on Hidden Markov Model (HMM)" International Journal of Advance Research in Computer Science and Management Studies, Volume 2, Issue 11, November 2014, pp.- 465-469.

[10]. Zhang Youzhi "Research and Application of Hidden Markov Model in Data Mining" Second IITA International Conference on Geoscience and Remote Sensing, IEEE, 2010, pp.-459-462.

[11]. Ritu. Sharma, Shiv Kumar and Rohit Maheshwari, "Comparative Analysis of Classification Techniques in DataMining Using Different Datasets", International Journal of Computer Science and Mobile Computing IJCSMC, vol. 4, no. 12, pp. 125-134, December 2015.

[12]. Sobhan Sarkar, Atul Patel, Sarthak Madaan and Jhareswar Maiti, "Prediction of Occupational Accidents Using DecisionTree Approach", IEEE Annual India Conference (lNDICON), pp. 1-6, 2016.

[13]. Verma and Shikha Mehta, "A Comparative Study of Ensemble LearningMethods for Classification in Bioinformatics", IEEE 7th International Conference on Cloud Computing Data Science & Engineering - Confluence, pp. 155-158, 2017.

[14]. K. C. Giri, M. Patel, A. Sinhal and D. Gautam, "A Novel Paradigm of Melanoma Diagnosis Using Machine Learning and Information Theory", 2019 International Conference on Advances in Computing and Communication Engineering (ICACCE), pp. 1-7, 2019.

[15]. Ayisheshim Almaw and Kalyani Kadam, "Survey Paper on Crime Prediction using EnsembleApproach", International Journal of Pure and Applied Mathematics, vol. 118, no. 8, pp. 133-139, 2018.

[16]. Shakuntala Jatav and Vivek Sharma, "An Algorithm for Predictive DataMining Approach In Medical Diagnosis", International Journal of Computer Science & Information Technology (IJCSIT), vol. 10, no. 1, pp. 11-20, February 2018.

[17]. Han Wu, Shengqi Yang, Zhangqin Huang, Jian He and Xiaoyi Wang, "Type 2 diabetes mellitus prediction model based on data mining", ELSEVIER Informatics in Medicine Unlocked, pp. 100-107, 2018.

[18]. M. Patel, N. Choudhary, N. Modi, P. Verma and B. Trivedi, "Designing an Enhanced Simulation Module for Multimedia Transmission Over Wireless Standards", Proceedings of International Conference on Communication and Networks. Advances in Intelligent Systems and Computing, vol. 508, 2017.

[19]. Sumalatha and R Santhi, "A Study on Hidden Markov Model (HMM)", International Journal of Advance Research in Computer Science and Management Studies, vol. 2, no. 11, pp. 465-469, November 2014.

[20]. Janardhanan Padmavathi, L. Heena and Fathima Sabika, "Effectiveness Of Support Vector Machines In Medical Data Mining", Journal Of Communications Software And Systems, vol. 11, no. 1, pp. 25-30, March 2015.

Cite this article as :

Minchala Siva Krishna, Mr. K. Naveen, " Ensemble Based Heart Disease Prediction Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 4, pp.184-190, July-August-2023.