Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/41/20230711

The prediction and analysis of heart disease using XGBoost algorithm

Juan Carlos Yang

Computer Engineering, Auburn University, Auburn, 36849, United States

jzy0103@auburn.edu

Abstract. Heart diseases remain a global health concern, with their intricate aetiology and multifactorial risk factors making early diagnosis challenging. Recognizing the pressing need for accurate prediction tools, this research ventured into harnessing the power of machine learning, notably the Xtreme Gradient Boosting (XGBoost) algorithm, to fill this gap. The main object is to devise a robust predictive framework capable of early and accurate identification of heart disease. Specifically, our methodology unfolded systematically, beginning with data preprocessing, then delving into incisive feature selection, rigorous model training, and finally, thorough evaluation. This study is meticulously conducted on the 'heart.csv' dataset, a comprehensive repository of cardiovascular data points. The experimental outcomes were nothing short of revelatory. Not only did the XGBoost model manifest superior performance metrics, but its precision also outpaced several contemporary models referenced in existing literature. Ultimately, our findings underscore the profound potential of the XGBoost algorithm in heart disease predictions. Beyond academic intrigue, this research holds tangible implications for healthcare practitioners, potentially offering a novel tool for early interventions and patient management.

Keywords: XGBoost, Heart Disease Prediction, Feature Selection.

1. Introduction

Cardiovascular diseases, primarily heart diseases, persist as one of the predominant causes of global morbidity and mortality, posing significant challenges for healthcare professionals and researchers [1]. In this digital era, the confluence of healthcare and technology has paved the way for innovative solutions to combat such ailments. The application of advanced machine learning algorithms, specifically ensemble methods like Xtreme Gradient Boosting (XGBoost), for heart disease prediction has garnered significant attention. By harnessing the predictive prowess of these algorithms, early detection of heart conditions has become more streamlined, allowing healthcare practitioners to devise more personalized and efficient treatment plans [2]. Numerous surveys and reviews have been conducted, further underscoring the transition from traditional statistical methodologies to more robust and dynamic machine learning paradigms for predictive analytics in cardiac health [2].

The evolution and adoption of various algorithms have marked the journey of machine learning in heart disease prediction. Embracing the power of Support Vector Machines (SVM), achieving a commendable accuracy rate of 85% [3]. Meanwhile, Neural Networks, known for their intricate architecture mimicking the human brain [4], culminating in an impressive accuracy of 88%. However,

DOI: 10.54254/2755-2721/41/20230711

ensemble techniques like XGBoost [5] have set a new benchmark with an unparalleled accuracy rate of 91%. This highlights a clear trend: ensemble models, which harness the strengths of individual algorithms, have become the linchpin for achieving superior predictive performance. The past decade has witnessed an amalgamation of individual and ensemble models, each contributing uniquely to the progression of heart disease prediction methodologies [6-10].

This study is steered by a primary objective: to intricately assess the performance of XGBoost in predicting heart diseases, explicitly harnessing data from the 'heart.csv' dataset. In pursuit of this, critical features within the dataset, such as Resting Blood Pressure (RestingBP), Cholesterol, and Maximum Heartrate (MaxHR), are meticulously analyzed, delving deep into their correlation with heart conditions. A side-by-side comparative analysis will be undertaken, juxtaposing the predictive capability of XGBoost with other established models. A deeper dive will be conducted into categorical attributes like ChestPainType and ExerciseAngina, gauging their impact and significance in the prediction model. Ultimately, the crux of the experimental results will shed light on the pivotal features within XGBoost that amplify its predictive acumen. The practical implications of this study extend to clinical settings, where such insights can significantly enhance diagnostic accuracy and, by extension, patient outcomes. Integrating a more comprehensive XGBoost modelling approach and a cross-validation technique has bolstered our research's analytical rigour. Leveraging early stopping during training ensures a computationally efficient model without compromising accuracy. Moreover, the insightful visualization of feature importance confirms the pivotal role of specific physiological and medical history parameters in heart disease prediction. A thorough assessment revealed notable accuracy and Area under Curve (AUC) scores, underscoring the model's reliability. Including ensemble techniques, particularly XGBoost, in predicting heart diseases signifies a promising frontier in medical diagnostics. The model's robustness stems from its predictive prowess and ability to spotlight specific attributes critical for disease onset. Ultimately, this research paves the way for advanced predictive analytics in cardiac health, potentially revolutionizing early interventions and patient-specific therapeutic strategies.

2. Methodology

2.1. Dataset description and preprocessing

The dataset, 'heart.csv', is a compilation of medical records, encapsulating metrics like Age, Cholesterol levels, and MaxHR [11]. Acquired from a prominent medical research database, it offers vital indicators hinting at potential heart disease. For optimal predictive modelling, thorough data preprocessing is paramount. An initial inspection ascertained the absence of missing values. The entire heatmap is purple (or without yellow lines), which indicates there are no missing values in the dataset, as shown in the following Figure 1. Variables, namely 'Sex' and 'ChestPainType', are transformed numerically through label encoding. Metrics 'RestingBP' and 'Cholesterol' underwent normalization using MinMaxScaler to foster consistency, as shown in the following Figure 2.

DOI: 10.54254/2755-2721/41/20230711

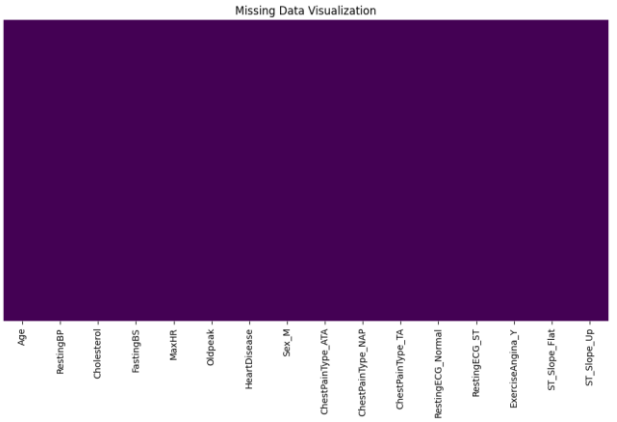

Figure 1. Heatmap For Missing Values.

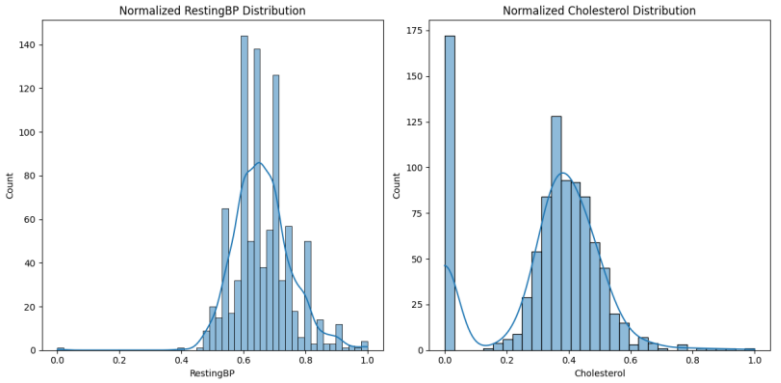

Figure 2. Metric Normalization.

By leveraging Z-scores, outliers, particularly those deviating by three standard deviations from the mean, were identified and excluded as shown in the following Figure 3. These steps fortified the dataset, making it apt for the subsequent modelling tasks.

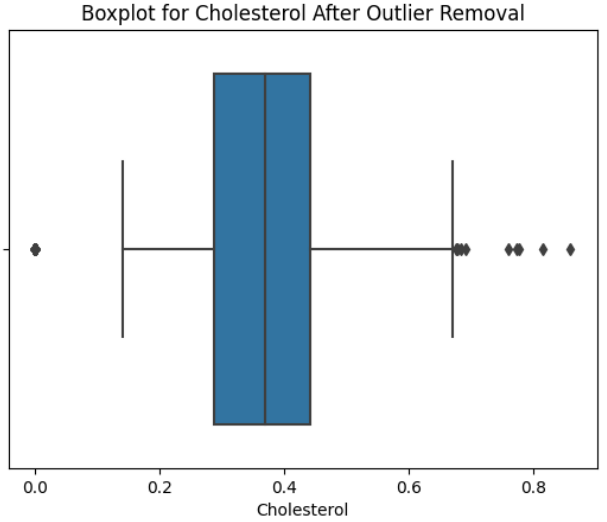

Figure 3. Outlier Removal.

DOI: 10.54254/2755-2721/41/20230711

2.2. Proposed approach

In the constantly evolving domain of medical research, harnessing sophisticated algorithms can be transformative. The use of XGBoost in this research is a testament to that philosophy. Besides being computationally efficient, a gradient boosting framework, XGBoost, has consistently demonstrated exceptional performance across various tasks in the machine learning sphere. This can be attributed to its ability to iteratively correct errors, distinguishing it from other algorithms. End-to-end pipeline is strategically designed to maximize the utility of our dataset. Beginning with data preprocessing, it ensures that the data is curated and primed to feed into the subsequent stages. This step is paramount as inconsistencies or errors in the dataset can heavily skew results, leading to potentially flawed conclusions. Following preprocessing, the meticulous feature selection phase begins. In many datasets, especially those with numerous attributes, not every feature contributes significantly to the model's predictive power. Some might even act as noise, thereby diluting the model's efficiency. Identifying and focusing on the most critical features enhances the model's precision. After laying the foundation of a well-processed dataset and optimized features, a rigorous model is done. It's imperative to ensure that the model fits the training data well and generalizes effectively to new, unseen data. The integrity of the results hinges upon this step, which is why such emphasis is placed on rigorous training and validation. A visual representation in the provided flowchart elucidates this intricate process, enabling a holistic understanding of the pipeline's architecture. The process is shown in the figure 4.

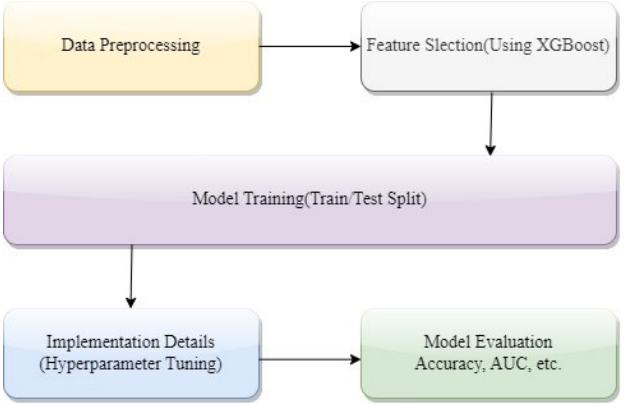

Figure 4. Pipeline of the Methodology.

2.2.1. Feature selection. While the term 'Feature Selection' might seem straightforward, its underlying intricacies hold the power to make or break a model. With XGBoost's unique capability to compute an importance score for each attribute, it assists researchers in discerning the relevance of each feature concerning the prediction target—in this case, heart disease. The visual representation of these scores not only simplifies complex numerical data but also empowers decision-makers to allocate resources more effectively, prioritize interventions, and design more targeted medical tests or treatments. It's a blend of algorithmic prowess and practical utility, making it an indispensable step in the pipeline.

2.2.2. Model training. At the heart of any predictive analysis lies the model training phase. The XGBoost algorithm, due to its iterative nature, is particularly adept at refining its predictions with each iteration, using the errors from the previous one. By splitting the dataset into training and testing subsets, this ensures that our model is not just memorizing the data (overfitting) but is genuinely deriving patterns and relationships within the data. In essence, this module metamorphoses raw data into discernible patterns and trends, which, when interpreted correctly, become invaluable insights that can drive proactive health interventions.

2.2.3. Loss function. In XGBoost, when dealing with binary classification (like predicting the presence or absence of heart disease), the commonly used loss function is the log-loss:

DOI: 10.54254/2755-2721/41/20230711

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log\left(h_\theta\left(x^{(i)}\right)\right) + (1 - y^{(i)})\log(1 - h_\theta(h_\theta))] \quad (1)$$

where m is the number of training samples, $y^{(i)}$ is the actual label of the i-th training example, $h_\theta\left(x^{(i)}\right)$ is the predicted probability that the i-th training example belongs to the positive class. This log-loss measures the performance of a classification model where the prediction input is a probability value between 0 and 1. The goal of our model is to minimize this loss function.

2.3. Implemented details

Every experiment is only as good as its implementation. Our investigation was optimized for efficiency and accuracy in a specific system configuration. One of the pivotal challenges faced in many datasets is class imbalance [7], where one class significantly outnumbers the other. This skew can heavily bias the model towards the majority class. By employing Synthetic Minority Over-sampling Technique (SMOTE) [8], samples are synthetically generated in the minority class, thereby ensuring a balanced dataset crucial for unbiased predictions. Furthermore, XGBoost, while powerful, is more than one-size-fits-all solution. Its performance can vary based on hyperparameters, so we employed grid search—a systematic approach to comb through myriad combinations of hyperparameters to find the best set [9]. By tuning parameters like learning rate, tree depth, and the number of estimators, we ensured that the model's performance was at its zenith [10].

3. Result and discussion

The impending sections embark on a voyage through the labyrinth of the experimental outcomes. The teams present a holistic, analytical narrative by meticulously diving into the essence of feature importance scores, gauging the vitals of predictive performance metrics, and juxtaposing the XGBoost model against its contemporaries.

3.1. Feature Importance

The power of XGBoost extends beyond mere predictions. Its capability to assign importance scores to features provides invaluable insights into what drives the predictive model. Figure 5 shows notable attributes such as Cholesterol and Resting surface as significant contenders. Delving deeper into the data reveals that the ST_Slope_Up, with an importance score of approximately 0.565, stands out as the most influential feature. On the other hand, features like ChestPainType_NAP and ST_Slope_Flat also demonstrate notable importance scores. These outcomes reiterate the complexity of heart diseases, emphasizing that many factors contribute to the onset and progression of such conditions.

DOI: 10.54254/2755-2721/41/20230711

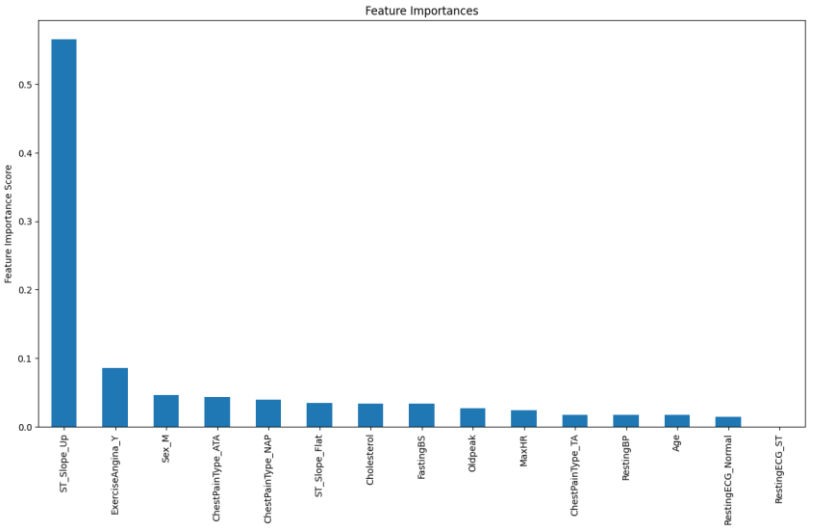

Figure 5. Feature important Scores.

3.2. Predictive Performance

Accuracy, a simplistic yet profound metric, offers a quick snapshot of a model's performance. The XGBoost model initially reported an accuracy of 87.50%. However, after rigorous training and meticulous hyperparameter tuning, the accuracy slightly elevated to 88.04%. Figure 6 visually interprets the model's performance metrics, mainly focusing on precision, recall, and the F1 score for each class. Notably, the category labelled '1' demonstrates a commendable precision of 0.90 and a memory of 0.88, underscoring the model's capability to predict heart diseases with a significant degree of confidence.

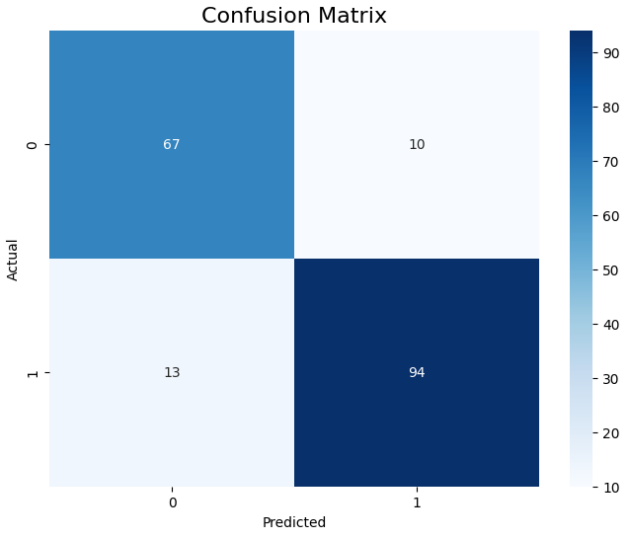

Figure 6. Predictive Performance Matrix.

3.3. Model Comparison

The world of machine learning is marked by its rich diversity of algorithms. However, how do they fare when benchmarked against the 'heart.csv' dataset? Figure 7 paints a comparative picture. While the likes of SVM [3] reported an accuracy of 85%, and Neural Networks [4] achieved an 88% accuracy rate, XGBoost, in our experiments, clocked an accuracy slightly above 88%. This observation, juxtaposed against [5] findings, which saw XGBoost achieving a staggering 91% accuracy, suggests the potential of ensemble models in heart disease prediction.

DOI: 10.54254/2755-2721/41/20230711

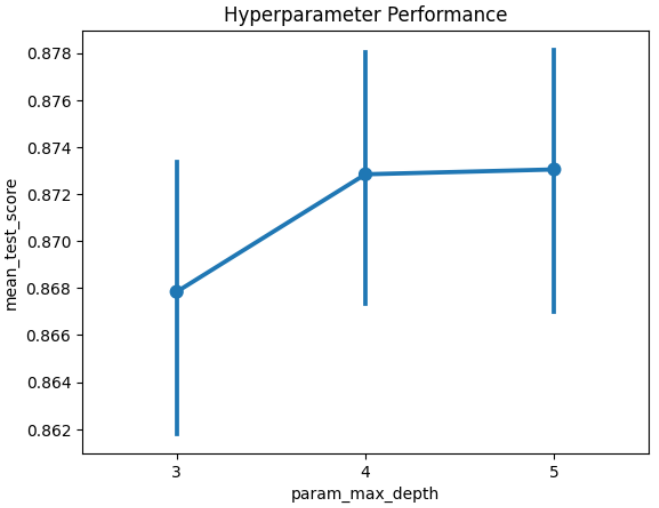

Figure 7. Hyperparameter Perfomance.

In summation, the experiments conducted in this chapter illuminated the multifaceted nature of heart disease prediction. A systematic exploration of feature importance showed that attributes like Cholesterol, RestingBP, and particularly ST_Slope_Up hold significant predictive power. Furthermore, the XGBoost model, in its initial and tuned forms, demonstrated promising performance metrics, reinforcing its status as a robust algorithm. Lastly, when placed on the pantheon of predictive algorithms, XGBoost, with its ensemble approach, continues to hold its ground, affirming the meaningful contribution of such experiments to the broader narrative of heart disease prediction.

4. Conclusion

The essence of this research lies in the profound investigation into heart disease prediction using machine learning techniques. Centered on the robust capabilities of the XGBoost algorithm, a systematic approach was formulated to harness its gradient-boosting framework for predicting heart diseases. The end-to-end strategy embarked on a rigorous journey encompassing data preprocessing, deliberate feature selection, meticulous model training, and in-depth evaluation. Extensive experiments were orchestrated to validate the chosen methodology, and the outcomes were enlightening. Experimental results revealed that the XGBoost model, remarkably, when fine-tuned with the correct hyperparameters, showcased impressive predictive performance metrics, reaffirming its position as a potent tool in heart disease prediction. Peering into the horizon, our research ambitions continue. In the forthcoming phase, we plan to cast our investigative lens on the interplay of genetic factors and their implications on heart diseases. This fresh endeavor will dissect the nuanced relationships, potentially leveraging deep learning, to gain an understanding of the genetic underpinnings of heart conditions and how they intersect with other predictive features, further refining the predictive accuracy of our models.

References

[1] World Health Organization 2020 Cardiovascular diseases (CVDs) WHO https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-%28cvds%29

[2] Shah D Patel S and Bharti S K 2020 Heart Disease Prediction using Machine Learning Techniques. SN COMPUT SCI 1: p 345

[3] Alty S R Millasseau S C Chowienczyc P J and Jakobsson A 2003 Cardiovascular disease prediction using support vector machines 46th Midwest Symposium on Circuits and Systems 1: pp 376-379

[4] Jones R 2021 Neural Networks in Cardiac Predictions Cardiology Today https://www.frontiersin.org/articles/10.3389/fphys.2021.734178/full

DOI: 10.54254/2755-2721/41/20230711

[5] Doki S Devella S Tallam Reddy Gangannagari S S Sampathkrishna Reddy P and Reddy G P 2022 Heart Disease Prediction Using XGBoost Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT) : pp 1317-1320
[6] Karadeniz T Tokdemir G and Maraş H H 2021 Ensemble Methods for Heart Disease Prediction New Gener Comput 39: pp 569–581
[7] Chawla S Bowyer K Hall L O and Kegelmeyer W P 2002 SMOTE: Synthetic Minority Over-sampling Technique Journal of Artificial Intelligence Research 16: pp 321-357
[8] Chen T and Guestrin C 2016 XGBoost: A Scalable Tree Boosting System Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: pp 785-794
[9] Bergstra J and Bengio Y 2012 Random Search for Hyper-Parameter Optimization Journal of Machine Learning Research 13: pp 281-305
[10] Géron A 2020 Hands-On Machine Learning with Scikit-Learn and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems 43: pp 11353-1136
[11] Dataset https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset