**DEEP ENSEMBLE LEARNING FOR CARDIOVASCULAR DISEASE PREDICTION**

Uday Hese, Harsh Mane, Nimish Marathe, Siddesh Mundhe

*Department of Computer Science and Engineering (Artificial Intelligence), Vishwakarma Institute of Information Technology, Pune, Maharashtra*

**ABSTRACT:**

Machine learning and deep learning have emerged as transformative tools in addressing the global challenge of cardiovascular disease (CVD), the leading cause of mortality worldwide. This study focuses on leveraging ensemble techniques within the context of heart disease prediction, utilizing big data analytics and medical expertise to create precise predictive models. Ensemble methods, including Decision Trees, Adaptive Boosting, Bagging, Stacking, and Random Forests, were employed to mitigate overfitting and enhance accuracy by amalgamating predictions from diverse models. The results showcase significant improvements in predictive performance, with the Random Forest Ensemble Technique achieving an impressive accuracy of 99.70%. This approach not only enhances early detection but also enables proactive prevention strategies and tailored interventions, thereby reducing the overall cost of CVD treatment. The collaborative synergy between researchers, data scientists, and healthcare practitioners drives continuous innovation, leading to a paradigm shift in cardiovascular healthcare towards optimized patient outcomes and improved global health.

**KEYWORDS:**

*Ensemble learning, Heart disease, Machine learning, boosting, bagging, adaboost, stacking, Random Forest.*

**INTRODUCTION:**

The domain of cardiovascular disease (CVD) prediction has become increasingly critical due to its status as the leading cause of mortality worldwide. Despite advancements in healthcare, challenges persist in accurately predicting and preventing CVD-related incidents. The existing system often relies on traditional risk factor assessments, which may not capture the intricate nuances of individual patient profiles, leading to suboptimal outcomes and increased healthcare costs. Available methods, including single-model machine learning approaches, have shown promise but are limited in their ability to handle the complexity and variability inherent in CVD data.

The proposed system introduces a novel approach using deep ensemble learning for cardiovascular disease prediction. This system addresses several key challenges present in the current landscape. First, it leverages ensemble techniques such as Decision Trees, Adaptive Boosting, Bagging, Stacking, and Random Forests to amalgamate predictions from multiple models, enhancing predictive accuracy and robustness. Second, it incorporates big data analytics to process vast amounts of heterogeneous data sources, capturing subtle risk factors and improving the overall predictive power of the models.

The research contributes significantly to the field of CVD prediction in several ways. Firstly, it introduces a comprehensive ensemble learning framework that outperforms traditional single-model approaches, as evidenced by the significant improvement in predictive accuracy, with the Random Forest Ensemble Technique achieving an impressive 99.70%. Secondly, the system enables early detection of CVD risk, leading to proactive prevention strategies that can significantly reduce the overall cost of treatment and healthcare burden associated with CVD-related incidents. Thirdly, by leveraging deep learning techniques, the proposed system can discern nuanced correlations and patterns within the data, providing clinicians with more reliable predictive tools for personalized interventions. Fourthly, the collaborative ethos between researchers, data scientists, and healthcare practitioners fosters continuous innovation and drives the evolution of predictive modeling in cardiovascular healthcare. Lastly, the proposed system sets a new standard for precision and efficiency in CVD prediction, offering a beacon of hope for millions worldwide by optimizing patient outcomes and improving global health.

**LITERATURE REVIEW:**

We conducted research on predicting heart disease by combining Particle Swarm Optimization (PSO) with an ensemble classifier. Our proposed methodology employs PSO as a feature selection technique to eliminate the least important features. We then utilized ensemble methods as classifiers to reduce misclassification rates and enhance classification performance. Our experimental results demonstrate a significant improvement in learning accuracy by applying the Bagged Tree Ensemble Classifier on PSO-selected features. This model offers medical professionals a reliable tool for accurately predicting and diagnosing heart diseases early, using a subset of essential features. Moving forward, we aim to integrate Rough Sets with PSO and develop a Decision Support System (DSS) for early heart disease diagnosis.[1]

In this study, ensemble learning classification and prediction models were created to diagnose and classify the presence or absence of coronary heart disease in patient outcome predictions. Furthermore, the models were assessed based on their accuracies, sensitivities (or recalls), precisions, specificities, F-scores, Receiver Operating Characteristic (ROC) curves, Area Under the Curve (AUC), and Kolmogorov-Smirnov (K-S) measures. The developed classification and prediction models, utilizing the Adaptive Boosting algorithm, were ensemble learning classifiers characterized by high flexibility in adjusting a weighting vector to generate a robust single composite ensemble learning classification and prediction model through an optimally weighted majority vote from multiple weak classifiers.[2]

The primary aim of this research is to improve the accuracy of predicting heart disease in patients, which holds significant value for healthcare information systems. To address the challenge of unstructured data uncertainty, the study integrates a Fuzzy K-Nearest Neighbor (KNN) classifier with a Symbolic approach. The current findings highlight the success of employing an interval approach to transform data into symbolic form, resulting in enhanced system accuracy. Future enhancements could involve expanding the number of attributes within the existing system. Additionally, testing the symbolic Fuzzy K-NN classifier with unstructured data from healthcare industry databases, by converting it into fuzzified structured data with increased attributes and a larger dataset, can further boost the system's accuracy in patient prediction and diagnosis for heart disease.[3]

Heart disease, being inherently fatal, presents life-threatening complications such as heart attacks and death. Recognizing the significance of Data Mining in the Medical Domain, efforts are being made to apply pertinent techniques in disease prediction. Various research endeavors have investigated effective techniques employed by different researchers. Insights gleaned from previous work have informed the design of the proposed system architecture for this study. Although several classification techniques are commonly utilized for disease prediction, the Decision Tree classifier was chosen for its simplicity and accuracy. Various attribute selection measures, including Information Gain, Gain Ratio, Gini Index, and Distance measures, can be employed in this context.[4]

This study aims to enhance heart disease diagnosis compared to previous methods. We designed a heart disease prediction model to aid healthcare professionals in assessing patients' heart disease status using clinical data. Initially, we identified 14 crucial clinical features, including age, sex, chest pain type, blood pressure, cholesterol levels, fasting blood sugar, ECG results, heart rate, exercise-induced angina, old peak, slope, number of vessels, thalassemia type, and heart disease diagnosis. Subsequently, we built a prediction model using the J48 decision tree algorithm to classify heart disease based on these features, employing unpruned, pruned, and pruned with reduced error pruning techniques. Our findings reveal that the Pruned J48 Decision Tree with Reduced Error Pruning offers superior accuracy compared to the simple 3Pruned and Unpruned methods. Notably, fasting blood sugar emerged as the most critical attribute for classification, although it did not yield the highest accuracy.[5]

This study examines the predictive accuracy of heart disease using a combination of classifiers. Training and testing utilized the Cleveland heart dataset from the UCI machine learning repository. Various ensemble algorithms such as bagging, boosting, stacking, and majority voting were tested. Bagging improved accuracy by up to 6.92%, boosting by up to 5.94%, majority voting by up to 7.26%, and stacking by up to 6.93%. Comparison revealed that majority voting yielded the greatest accuracy enhancement. Additionally, the performance benefited from feature selection techniques, which further improved the ensemble algorithms' accuracy. The highest accuracy was achieved using majority voting with the FS2 feature set.[6]

The paper proposes expanding the model's attributes and dataset to broaden its scope and improve accuracy, suggesting the use of deep learning and ensemble techniques for further enhancement. It envisions real-time applications across various disorders, facilitating comparative analyses based on diseases and algorithms used. Connecting with wearable devices like smartwatches and fitness trackers, machine learning models could offer continuous monitoring and predictive insights, empowering individuals in managing their heart health proactively. To gain user trust, future research should focus on developing interpretable methods and transparent explanations for model predictions. This transparency not only boosts user confidence but also aids in making informed health decisions. Ensemble techniques play a crucial role in reducing overfitting and handling noisy data, making them more robust compared to single models. By aggregating predictions from diverse models, ensemble methods can filter out noise and capture underlying data patterns, leading to more accurate heart disease predictions. Continued innovation and collaboration hold promise for revolutionizing cardiovascular healthcare through predictive models, potentially advancing early detection, prevention, and personalized interventions for improved patient outcomes.[7]

The review indicates a significant potential for machine learning algorithms in predicting cardiovascular diseases. While various algorithms like Alternating Decision Trees with PCA have excelled in certain scenarios, others such as basic Decision Trees have shown limitations, possibly due to overfitting issues. Models like Random Forest and Ensemble models have fared well due to their ability to mitigate overfitting through multiple algorithm utilization. Naïve Bayes classifiers, known for their computational efficiency, also demonstrated good performance. SVM stood out for its consistent high performance across most cases. Although machine learning systems have shown high accuracy in predicting heart-related diseases, challenges like handling high-dimensional data and overfitting persist, warranting further research. Exploring the optimal ensemble of algorithms for specific data types remains an area ripe for investigation.[8]

Within the medical industry, predicting cardiovascular disease stands as a crucial area. This research delves into utilizing the patient's available data to forecast the presence or absence of cardiac issues. Various techniques exist for this prediction task, and here, the focus is on employing the Logistic Regression supervised Machine Learning algorithm. Enhancing performance involves pre-processing the data corpus through cleaning and addressing missing values. A pivotal aspect is feature selection, pivotal in boosting algorithm accuracy and understanding its behavior. Notably, as training progresses, Logistic Regression's accuracy in prediction also improves. In this study, the LR classifier achieved an accuracy of 87.10% with a training set of 90% and testing set of 10%, showcasing superior results compared to prior research. However, a limitation lies in using only the UCI dataset, prompting future work to expand onto multiple datasets for broader applicability.[9]

This paper introduces a novel integrated Bagging-Fuzzy-GBDT prediction algorithm aimed at improving heart disease diagnosis accuracy. Specifically, we incorporated fuzzy logic into GBDT to enhance its generalization ability. Additionally, integrating the Bagging algorithm with Fuzzy-GBDT helps prevent overfitting and facilitates data parallelization during training, leading to reduced training time. Simulation results demonstrate that our proposed Bagging-Fuzzy-GBDT algorithm exhibits significant improvements in accuracy, precision, AUC, and other metrics when compared to traditional algorithms. Combining the Bagging-Fuzzy-GBDT algorithm with IoT technology can enhance patient health monitoring and advance the convergence of IoT and machine learning in the medical domain. Future work will focus on optimizing the algorithm's complexity and training time to further enhance heart disease prediction performance.[10]

**METHODALOGY:**

ABOUT DATASET:

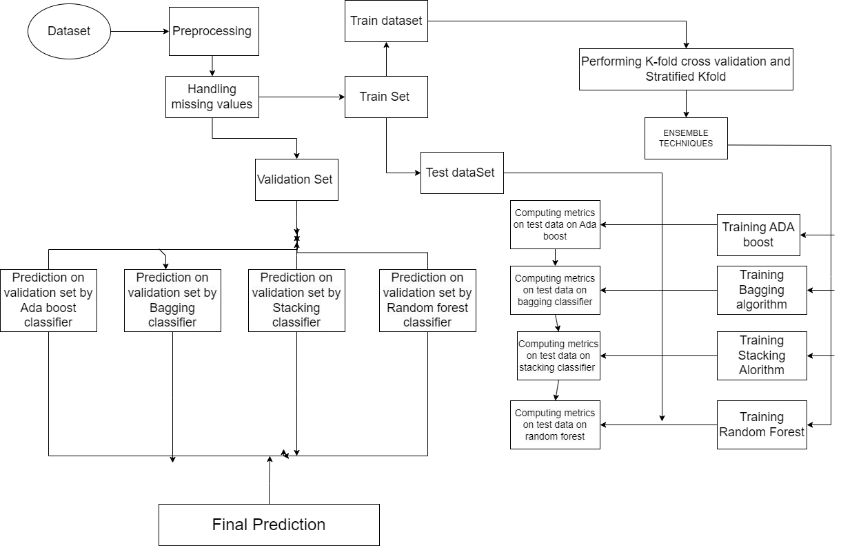
For this study, the Heart Disease Dataset was chosen as the primary dataset. This openly available dataset comprises information from four different sources: the Cleveland Clinic Foundation, Medical Centre Long Beach, Hungarian Institute of Cardiology, and University Hospital Switzerland.

The dataset contains 303 records, encompassing a total of 76 attributes. However, only 13 attributes and one target attribute were considered for this research. Table 1 provides a detailed overview of the attributes present in the dataset, comprising 8 categorical and 6 numeric attributes. These attributes encompass various clinical test results, including serum cholesterol levels, fasting blood sugar levels, vessel count, and thalassemia indicators derived from blood analysis. Additionally, ST depression and slope of ST-segment were derived from electrocardiogram data.

Attribute Description of the KAGGLE’s Heart Disease Dataset:

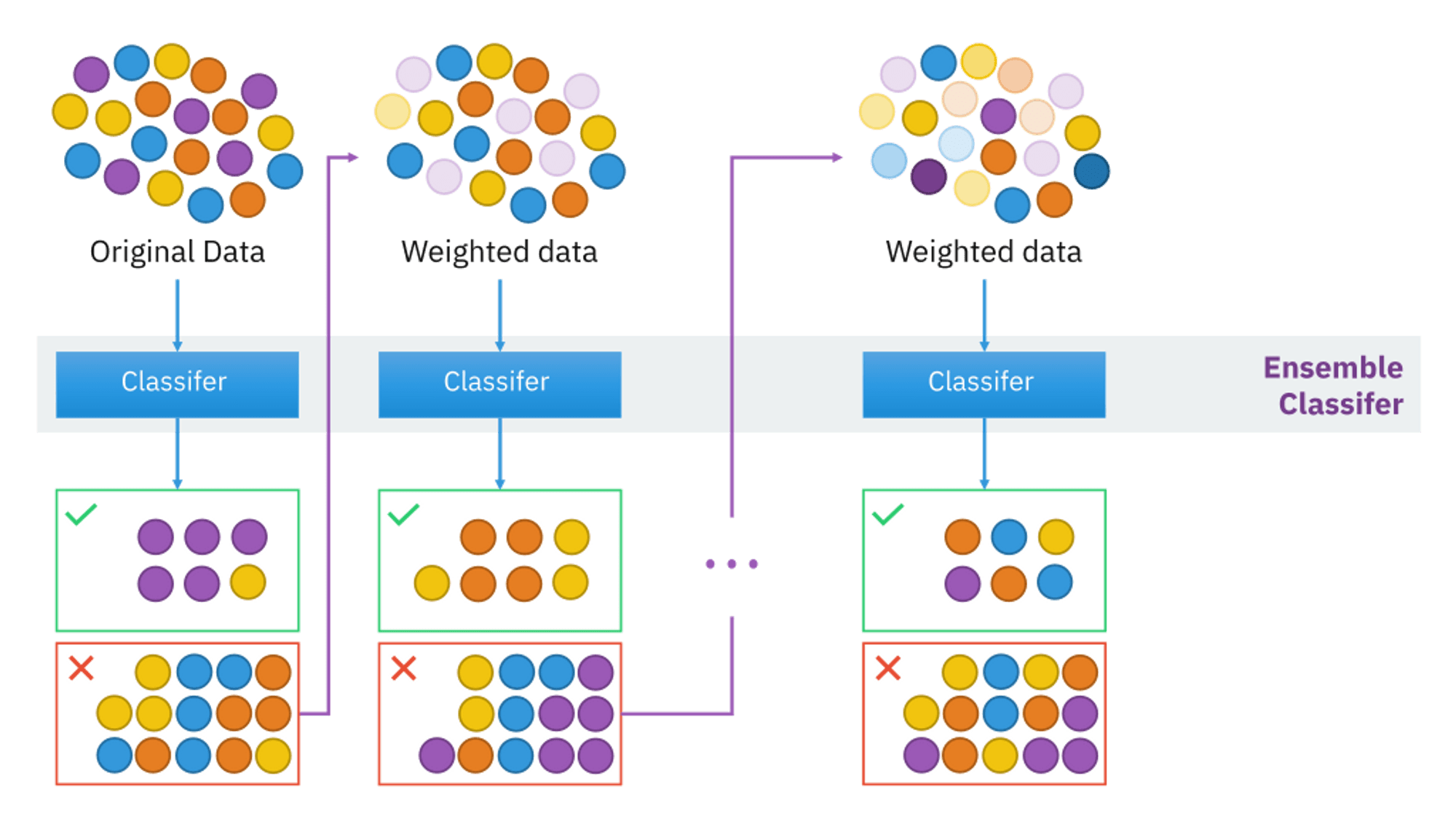
|  |  |
| --- | --- |
| **Attribute** | **Description** |
| Age | Age of the patient in years. |
| Sex | Gender of the patient (0 for Female, 1 for Male). |
| Chest Pain | Description of chest pain type (1 for Typical angina, 2 for Atypical angina, 3 for Non-anginal pain, 4 for Asymptomatic pain). |
| Resting Blood Pressure | Resting blood pressure measured in mm Hg. |
| Serum Cholesterol | Serum cholesterol level measured in mg/dl. |
| Fasting Blood Sugar | Fasting blood sugar level (>120 mg/dl indicated by 1, otherwise 0). |
| Rest Electrocardiograph | Resting electrocardiograph results (0 for Normal, 1 for ST-T wave abnormality, 2 for Left ventricular hypertrophy). |
| Maximum Heart Rate | Maximum heart rate achieved during exercise. |
| Exercise-induced Angina | Presence of exercise-induced angina (0 for No, 1 for Yes). |
| ST Depression | ST depression induced by exercise relative to rest. |
| Slope of ST Segment | Slope of the peak exercise ST segment (1 for upsloping, 2 for flat, 3 for downsloping). |
| Vessel Count | Number of major vessels colored by fluoroscopy (ranging from 0 to 3). |
| Thalassemia | Type of thalassemia (normal, fixed defect, reversible defect). |
| Heart Disease | Presence of heart disease (0 for negative, 1 for positive). |

FLOWCHART:



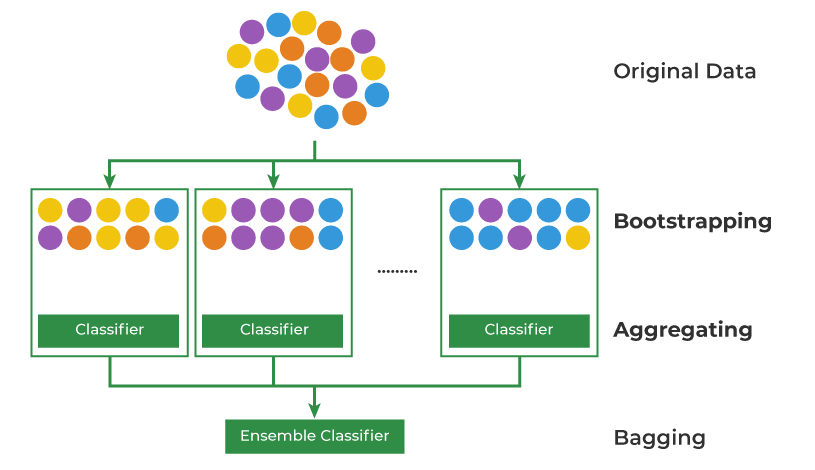
*Figure 1, (Architecture Diagram)*

1. **Adaptive Boosting Ensemble:** The AdaBoost algorithm, short for Adaptive Boosting, is a boosting technique used as an ensemble method in machine learning. This is called adaptive acceleration because the weights are reassigned for each case, with a higher weight for misclassified cases. This algorithm builds a model and assigns equal weight to all data points. It then assigns a higher weight to the misclassified points. Now, every focus with more weight gains more importance in the next model. It keeps the training patterns until a smaller error is obtained.



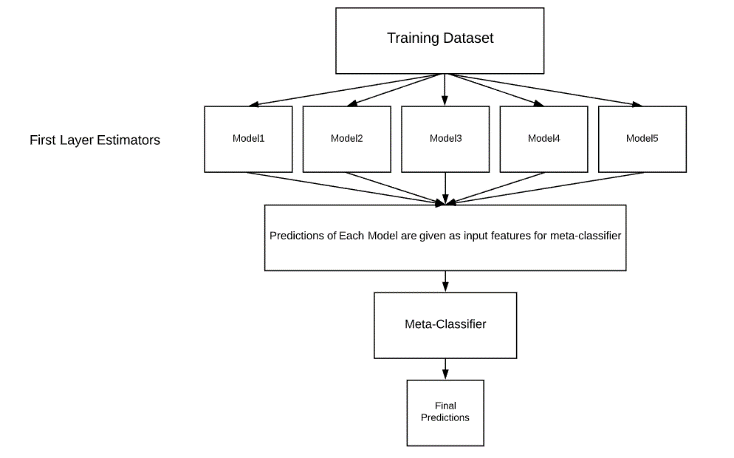
*Figure 2, (Ada Boost Working Diagram)*

1. **BAGGING Ensemble:** Bagging, also known as Bootstrap aggregation, is an ensemble learning technique that helps improve the efficiency and accuracy of machine learning algorithms. It is used to handle tradeoffs between bias and variance and reduces the variance of the predictive model. Bagging avoids data trophying and is used in both regression and classification models, especially decision tree algorithms. It is a homogeneous model of weak learners who learn independently of each other in parallel and combine them to determine the model average.



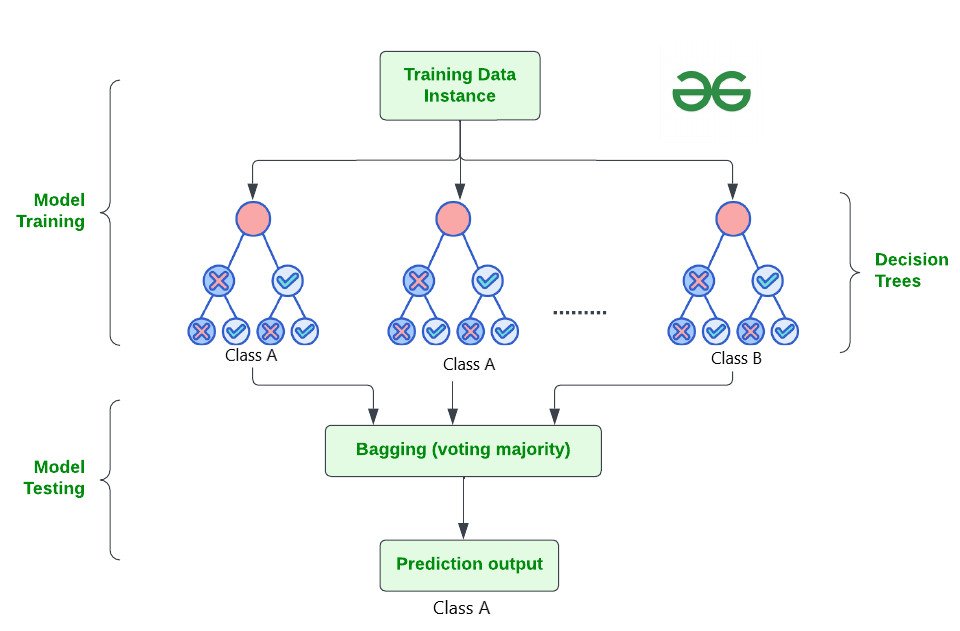
*Figure 2, (Bagging Ensemble Working Diagram)*

1. **STACKING Ensemble Technique:** Stacking (sometimes called stacked generalization) is a different paradigm. The purpose of stacking is to examine the state of different models for the same problem. The idea is that a learning problem can be attacked by different types of models that can learn part of the problem but not the entire state of the problem. This way you can build several different learners and use them to build an average prediction, one prediction for each learned model. Then you add a new model that learns from intermediate predictions of the same object. This final design is said to be layered on top of the others, hence the name. So you can improve your overall performance and often end up with a model that is better than any single average model.



*Figure 3, (Stacking Ensemble Working Diagram)*

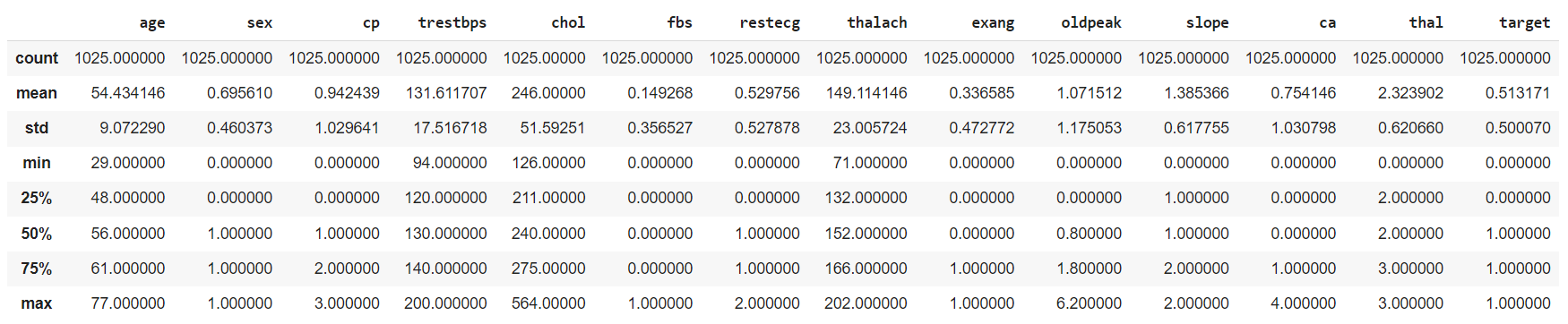
1. **Random Forest Ensemble**: The Random Forest algorithm is a powerful tree learning technique in machine learning. It works by generating multiple decision trees during the training phase. Each tree is constructed using a random subset of the dataset to measure a subset of random functions for each partition. This randomness introduces differences between individual trees, reducing the risk of overfitting and improving overall predictive power. In prediction, the algorithm aggregates the results of all the trees either by voting (classification tasks) or by calculating the average (regression tasks). This collaborative decision-making process, supported by multiple trees with their knowledge, is an example of stable and accurate results. . Random forests are widely used in classification and regression functions, known for their ability to handle complex data, reduce overfitting, and provide reliable predictions in various environments.



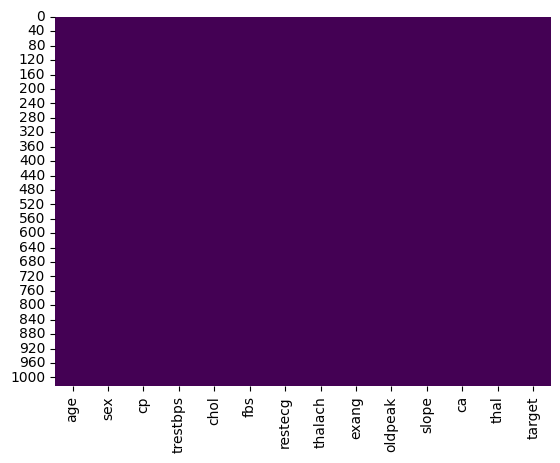
*Figure 3, (Random Forest Working Diagram)*

**RESULTS:**

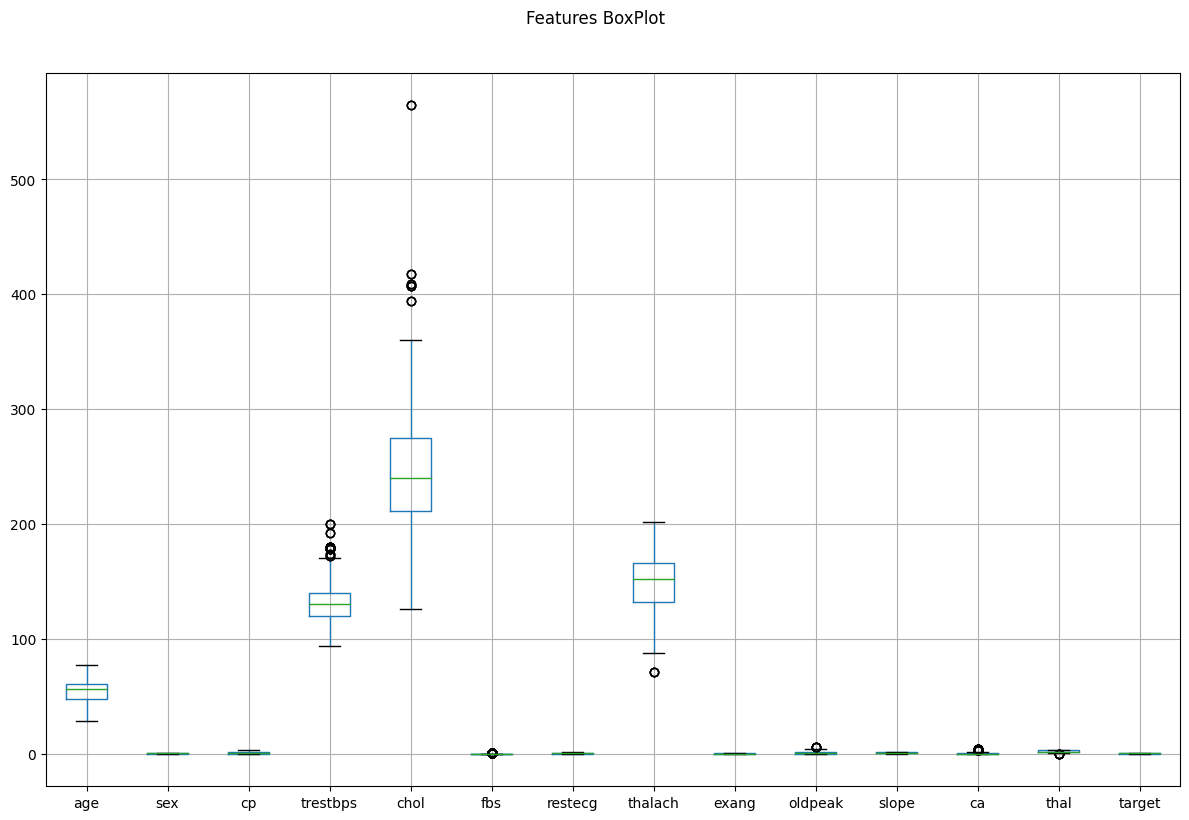
Exploratory Data Analysis (EDA):



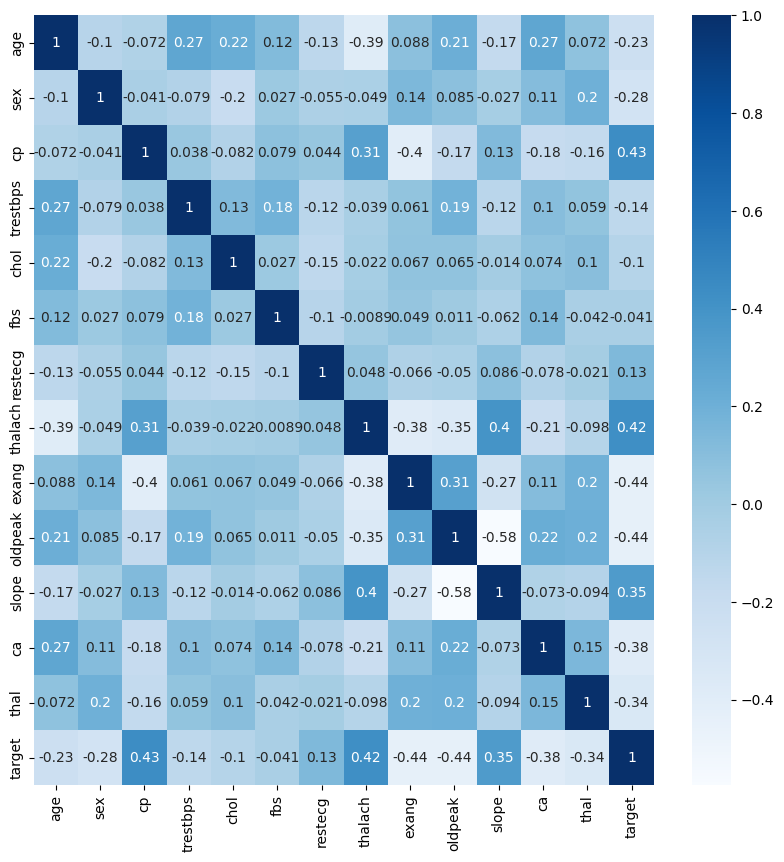
*Figure 6, (Data Exploration)*



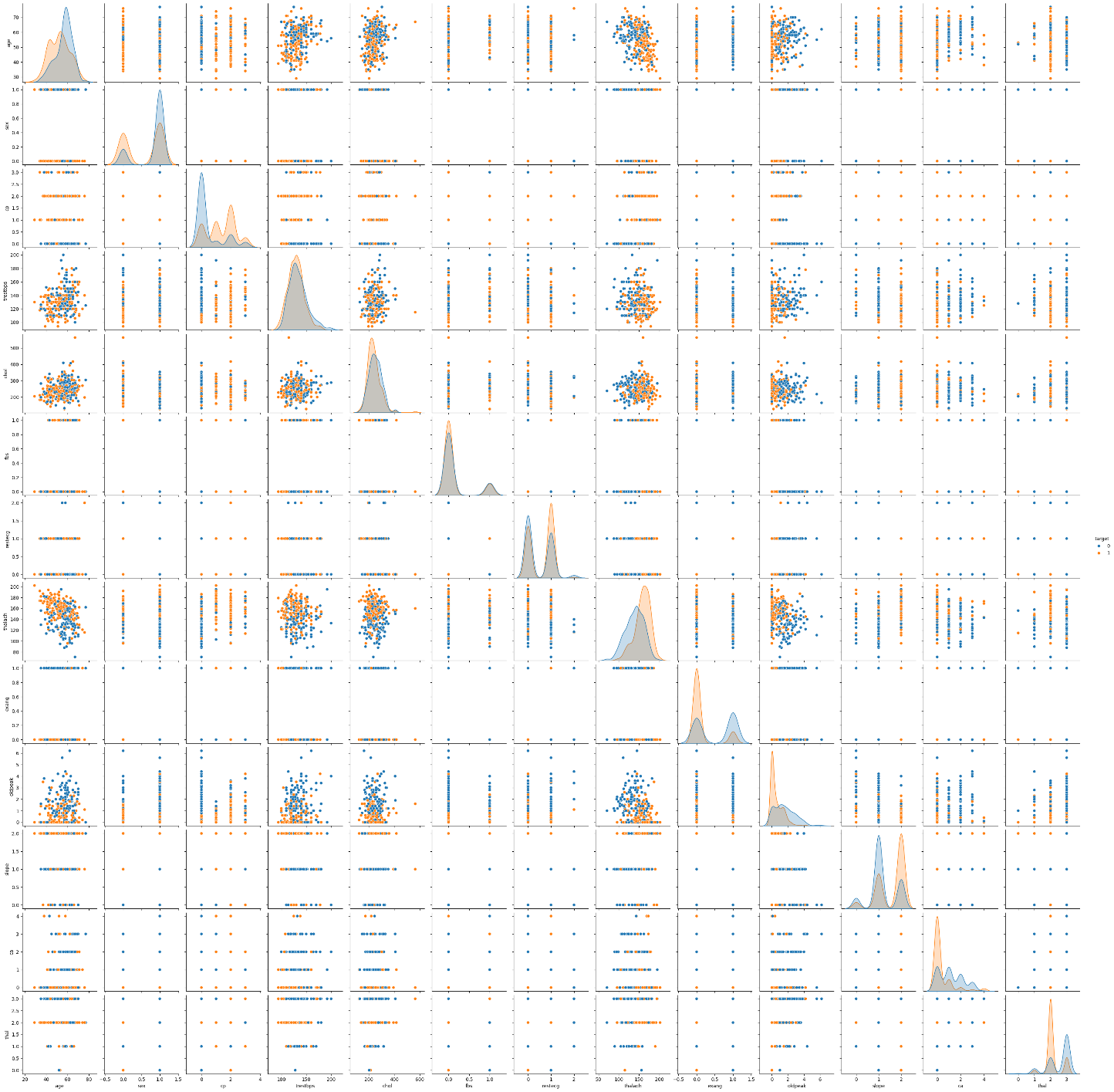
*Figure 7, (Visualizing the Missing Values With the help of heatmap)*



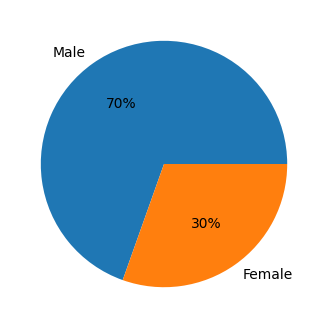
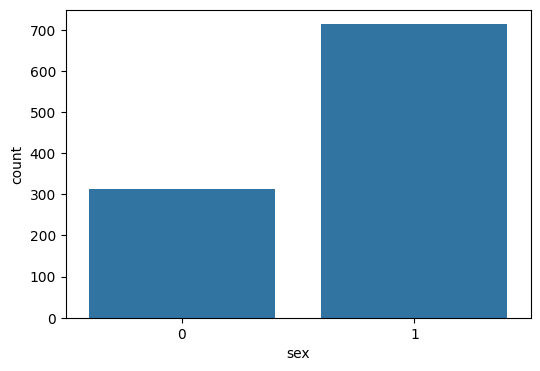
*Figure 8, (Boxplot for all features)*



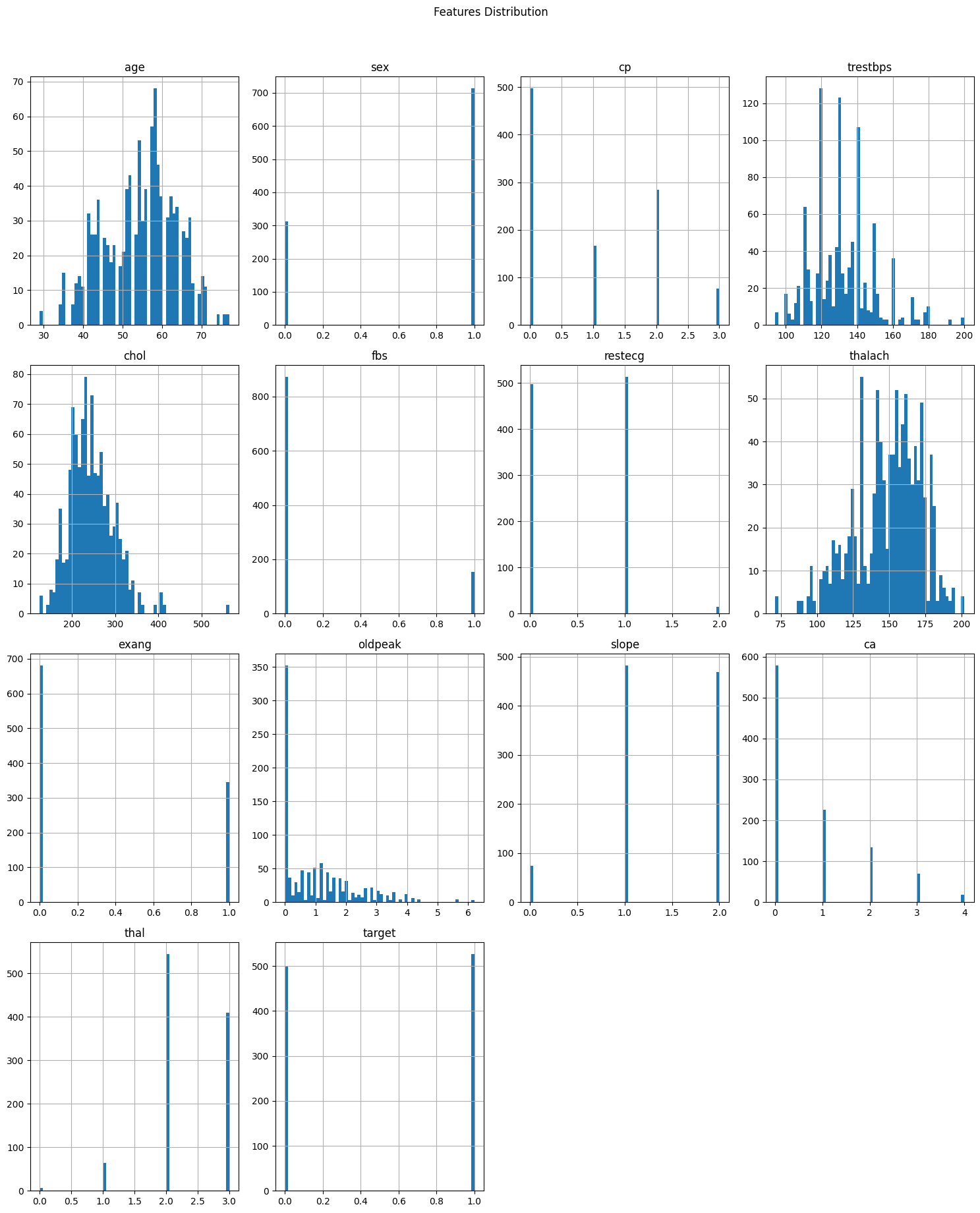
*Figure 9, (Correlation between different variables)*



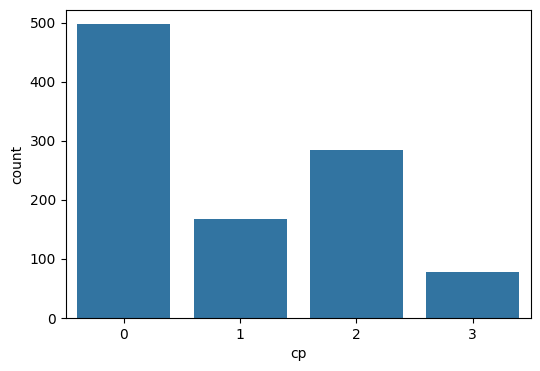
*Figure 10,(Pairplot for complete dataframe)*

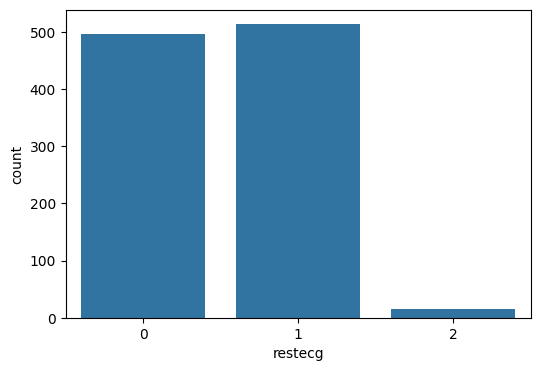
*Figure 11 , (Pie-Chart of Sex Column) Figure 12 , (Count Plot for Sex Column)*



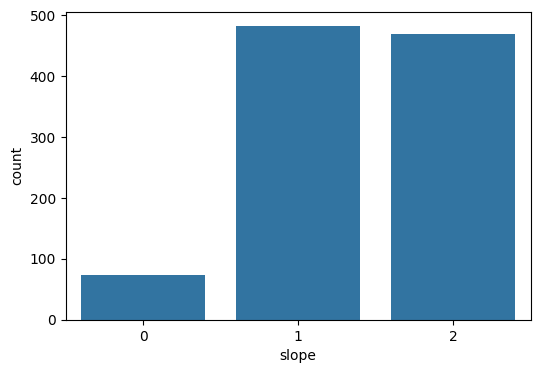
*Figure 13, (Features distribution histogram plot)*



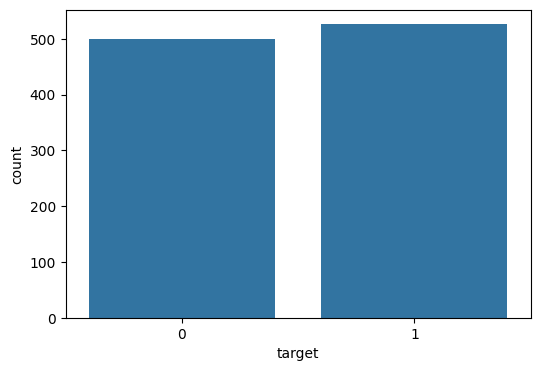
*Figure 14,(Count Plot for Chain Pain type column)*



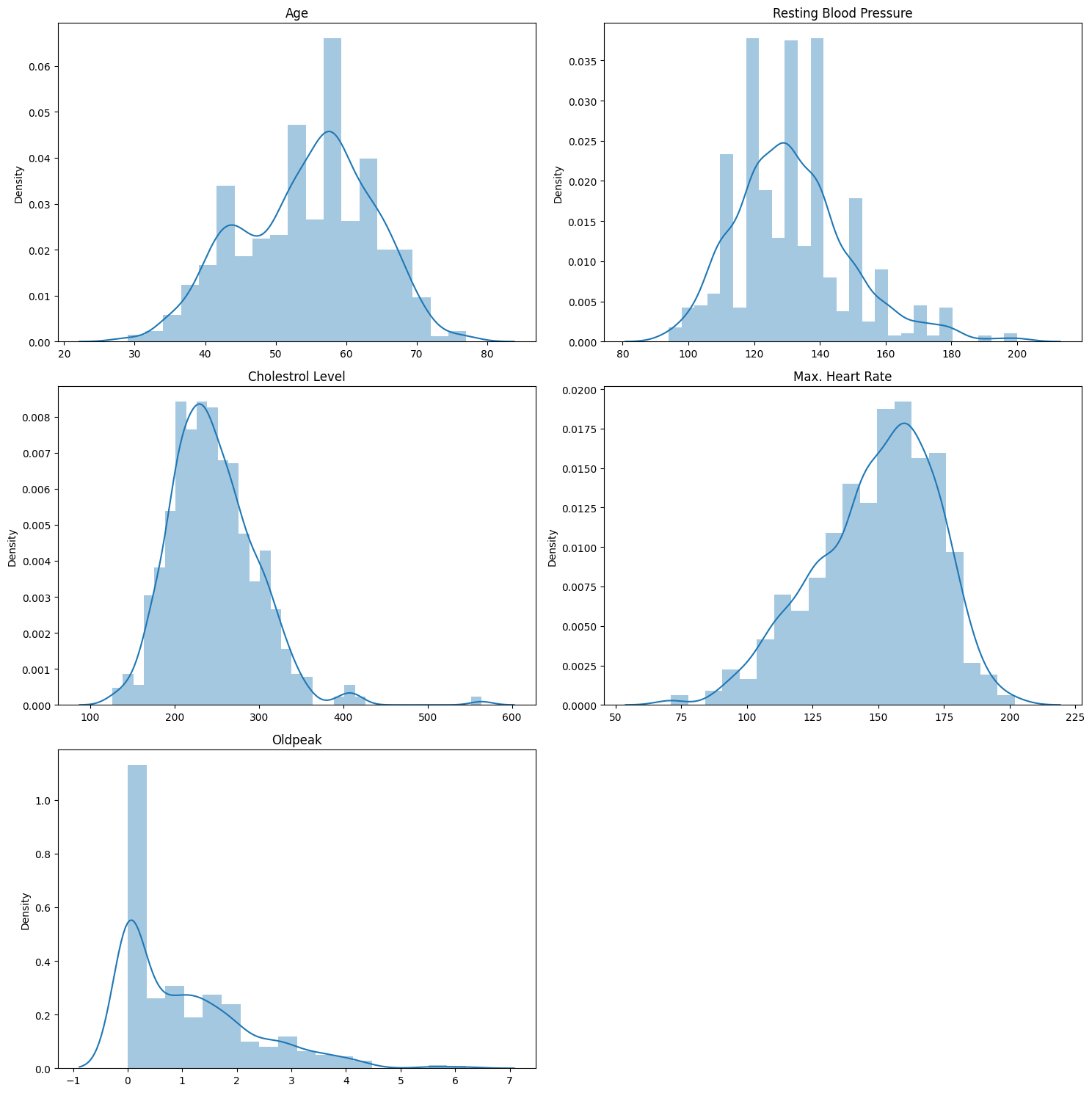
*Figure 15, (Count Plot for electrocardiographic results)*



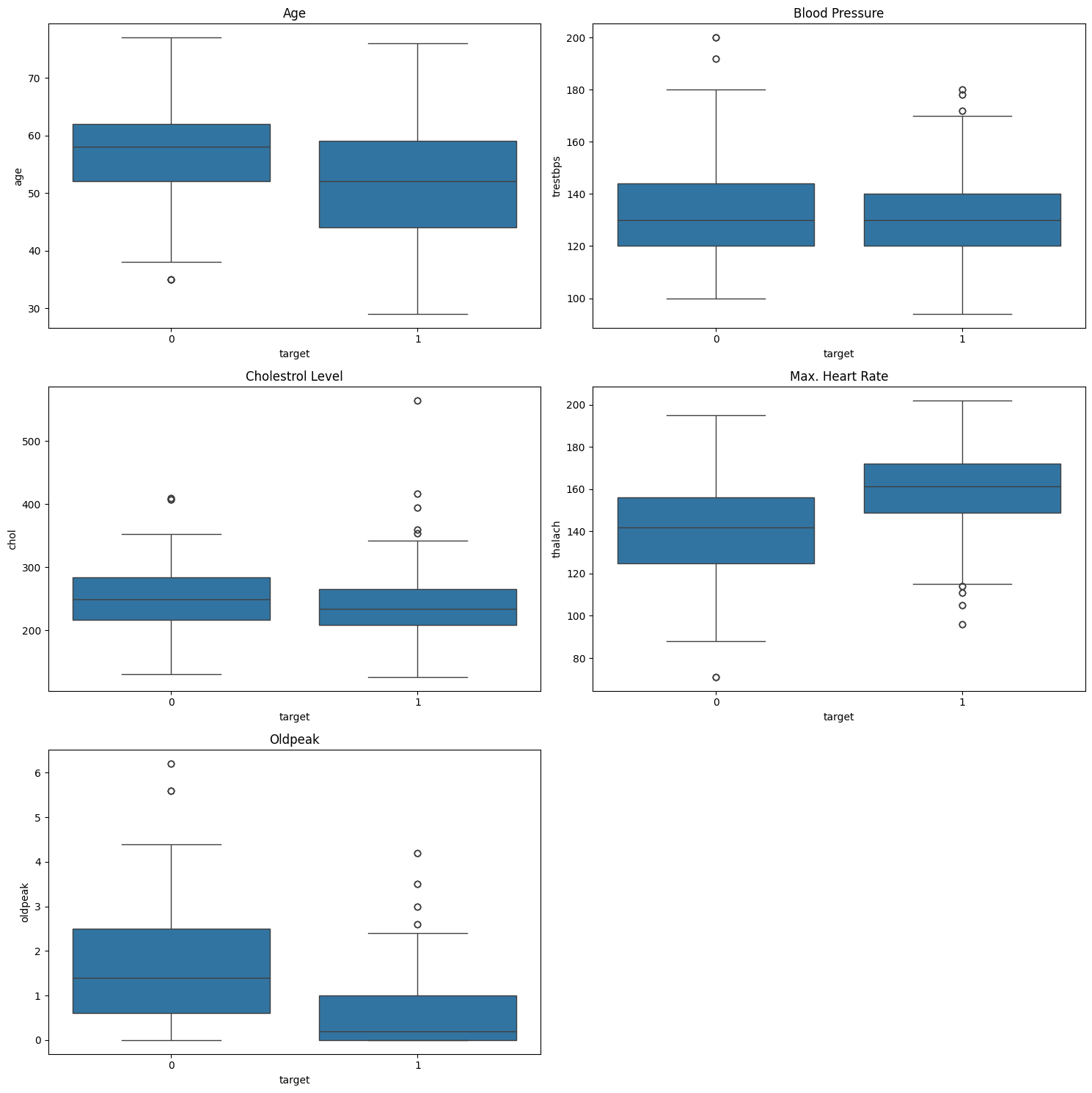
*Figure 16,* (*count plot for st slope column)*



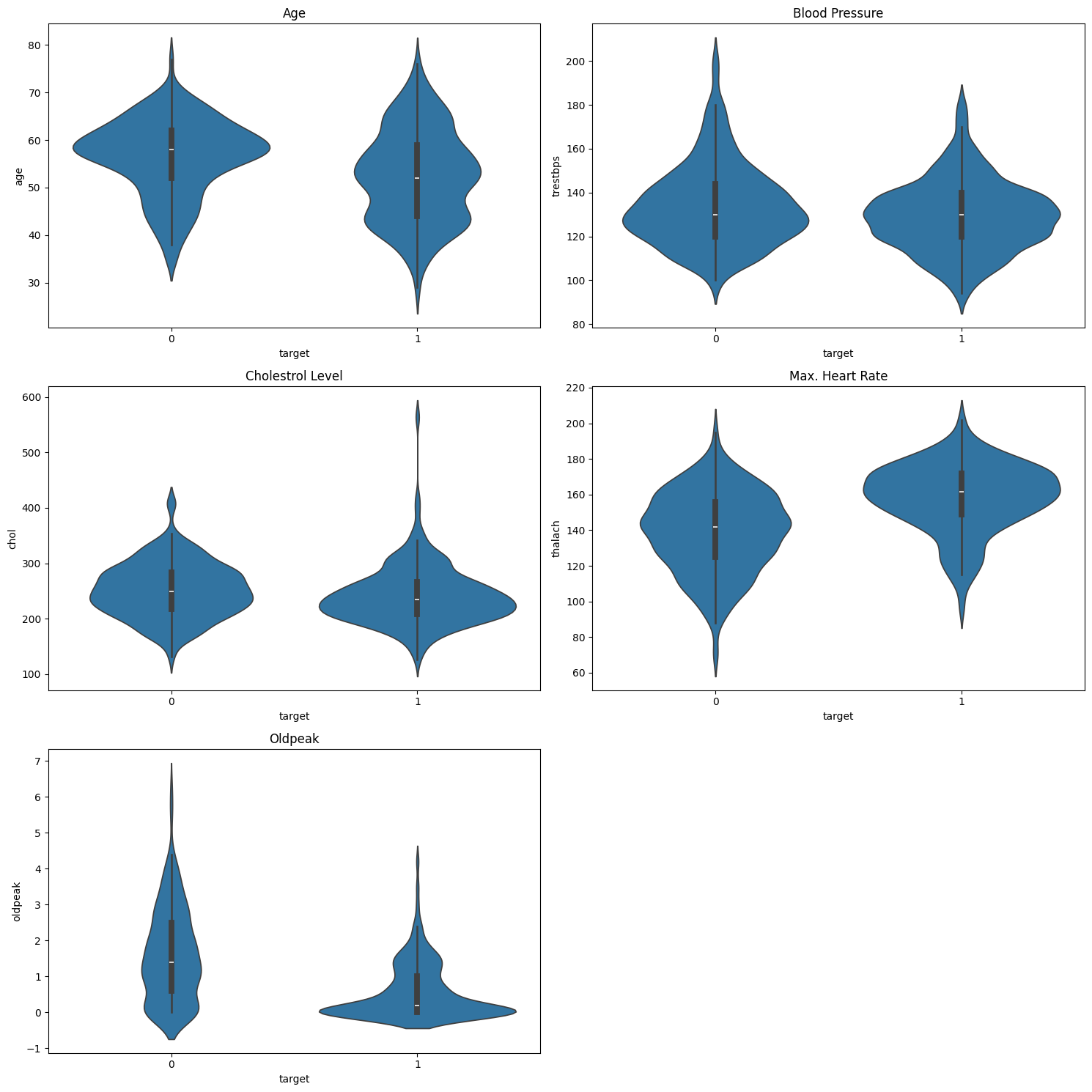
*Figure 17, (count plot for target column)*



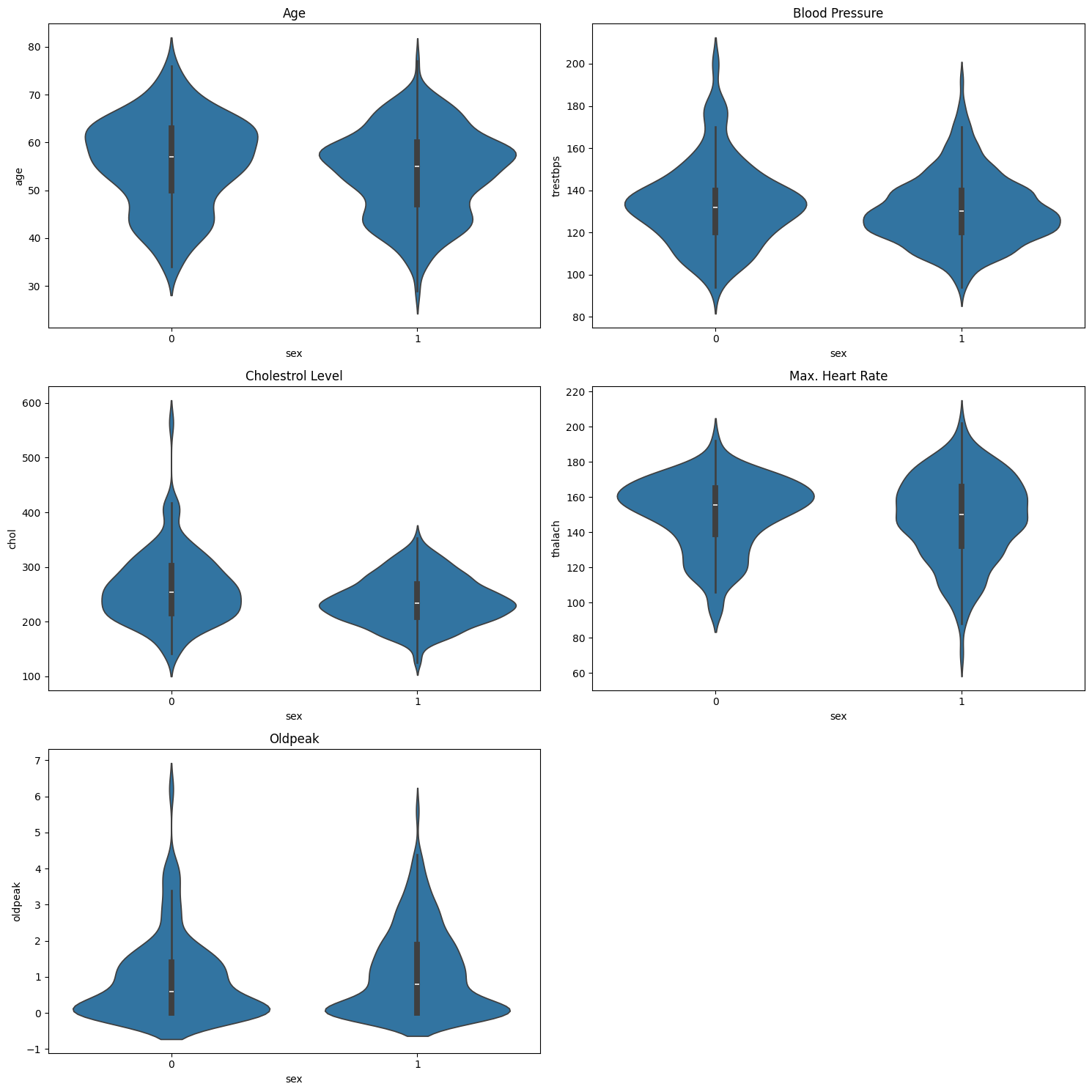
*Figure 18, (histogram with kde to see the distribution of features)*



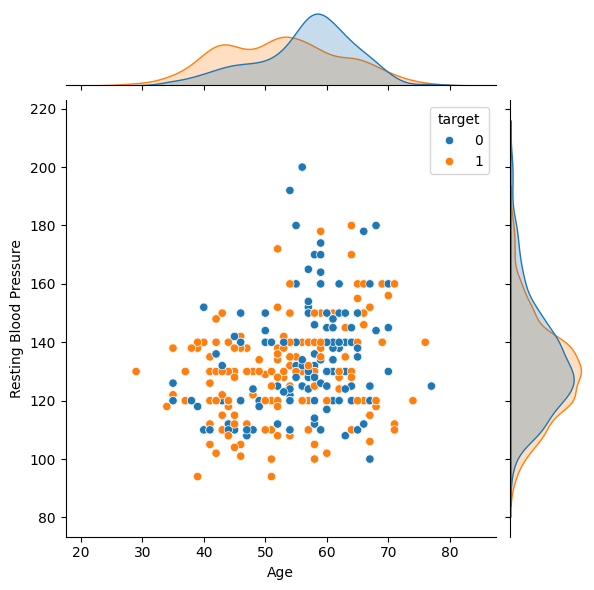
*Figure 19, (Boxplots to see the distribution)*



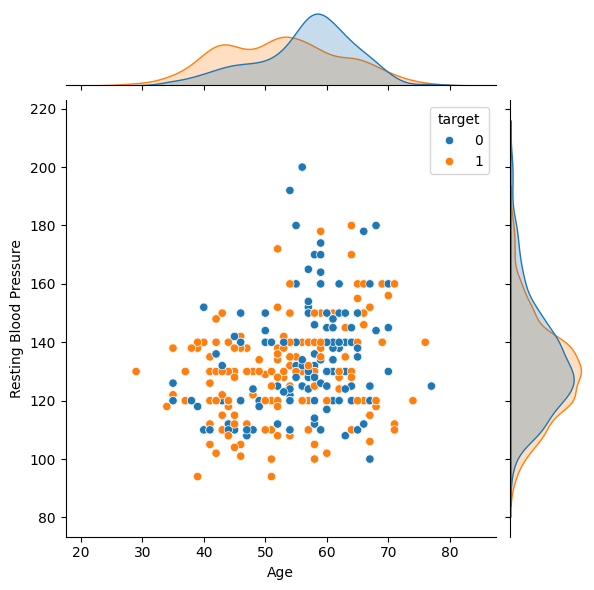
*Figure 20, (Violinplots for features ('Heart Disease Result' vs 'All numeric features')*



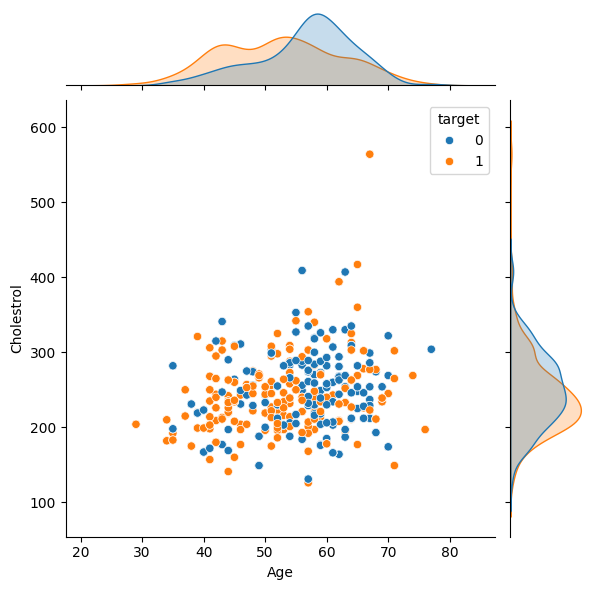
*Figure 21, (violinplots for features ('Sex' vs 'All numeric features'))*



*Figure 22, (Jointplot for 'Age' vs 'Max. Heart Rate')*



*Figure 23 , (Jointplot for 'Age' vs 'Blood Pressure')*



*Figure 24, (Jointplot for 'Age' vs 'Cholesterol')*

*Figure 25, (Accuracy Comparison)*

|  |  |  |
| --- | --- | --- |
| Sr. No. | ENSEMBLE TECHNIQUE /ML MODEL | ACCURACY |
| 1. | Decision Tree | 76% |
| 2. | Adaptive Boosting | 88.3% |
| 3. | Bagging Ensemble Technique | 98.44% |
| 4. | Stacking Ensemble Technique | 98.53% |
| 5. | Random Forest Ensemble Technique | 99.70% |

**CONCLUSION:**

This paper suggests that the model's scope can be varied by expanding its attributes and accuracy can be improved by expanding its dataset. Other algorithms like deep learning and ensemble techniques can be used to further enhance the model. It can be used in real time situations and could be applied to different disorders, enabling a comparative analysis of the model's performance based on diseases and algorithms utilized. Drawing connections with smartwatches and fitness trackers, ML models could provide continuous monitoring and predictive analysis empowering individuals to cautiously manage their heart health. The model still needs to gain users' trust thus future research could focus on developing methods to interpret model predictions and provide transparent explanations for decision-making. To ensure widespread acceptance and trust in the heart disease prediction model, it's essential to tackle concerns regarding how the model generates predictions. Users must comprehend the reasoning behind its decisions and feel assured in its accuracy. Hence, future research should prioritize developing techniques to interpret model predictions and offer transparent explanations for the influencing factors. This transparency not only boosts users' confidence in the model but also aids in making well-informed health decisions. Usage of the ensemble technique helps in reduced overfitting by averaging or combining the predictions of multiple models trained on different subsets of the data. Ensemble techniques are inherently more robust to noisy data compared to single models. By aggregating predictions from multiple models, ensemble methods can effectively filter out noise and focus on capturing the underlying patterns in the data, leading to more accurate heart disease predictions. Continued innovation and teamwork have the potential to revolutionize cardiovascular healthcare using predictive models. This could lead to advancements in early detection, prevention, and personalized interventions, ultimately enhancing patient outcomes.

**REFRENCES:**

[1] I. Yekkala, S. Dixit and M. A. Jabbar, "Prediction of heart disease using ensemble learning and Particle Swarm Optimization," 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon), Bengaluru, India, 2017, pp. 691-698, doi: 10.1109/SmartTechCon.2017.8358460.

[2] Miao, Kathleen H., Julia H. Miao, and George J. Miao. "Diagnosing coronary heart disease using ensemble machine learning." *International Journal of Advanced Computer Science and Applications* 7.10 (2016).

[3] V. Krishnaiah, M. Srinivas, G. Narsimha and N. S. Chandra, "Diagnosis of heart disease patients using fuzzy classification technique," International Conference on Computing and Communication Technologies, Hyderabad, India, 2014, pp. 1-7, doi: 10.1109/ICCCT2.2014.7066746.

[4] Thenmozhi, K., and P. Deepika. "Heart disease prediction using classification with different decision tree techniques." *International Journal of Engineering Research and General Science* 2.6 (2014): 6-11.

[5] Pandey, A. K., Pandey, P., Jaiswal, K. L., & Sen, A. K. (2013). A heart disease prediction model using decision tree. *IOSR Journal of Computer Engineering (IOSR-JCE)*, *12*(6), 83-86.

[6] Latha, C. Beulah Christalin, and S. Carolin Jeeva. "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques." *Informatics in Medicine Unlocked* 16 (2019): 100203.

[7] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[8] Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. *International Journal of Engineering & Technology*, *7*(2.8), 684-687.

[9] Ambrish, G., Ganesh, B., Ganesh, A., Srinivas, C., & Mensinkal, K. (2022). Logistic regression technique for prediction of cardiovascular disease. *Global Transitions Proceedings*, *3*(1), 127-130.

[10] X. Yuan et al., "A High Accuracy Integrated Bagging-Fuzzy-GBDT Prediction Algorithm for Heart Disease Diagnosis," 2019 IEEE/CIC International Conference on Communications in China (ICCC), Changchun, China, 2019, pp. 467-471, doi: 10.1109/ICCChina.2019.8855897.

[11] <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>

[12] <https://www.almabetter.com/bytes/tutorials/data-science/adaboost-algorithm>

[13] <https://www.geeksforgeeks.org/ml-bagging-classifier/>

[14] <https://www.geeksforgeeks.org/stacking-in-machine-learning-2/>

[15] <https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/>