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:

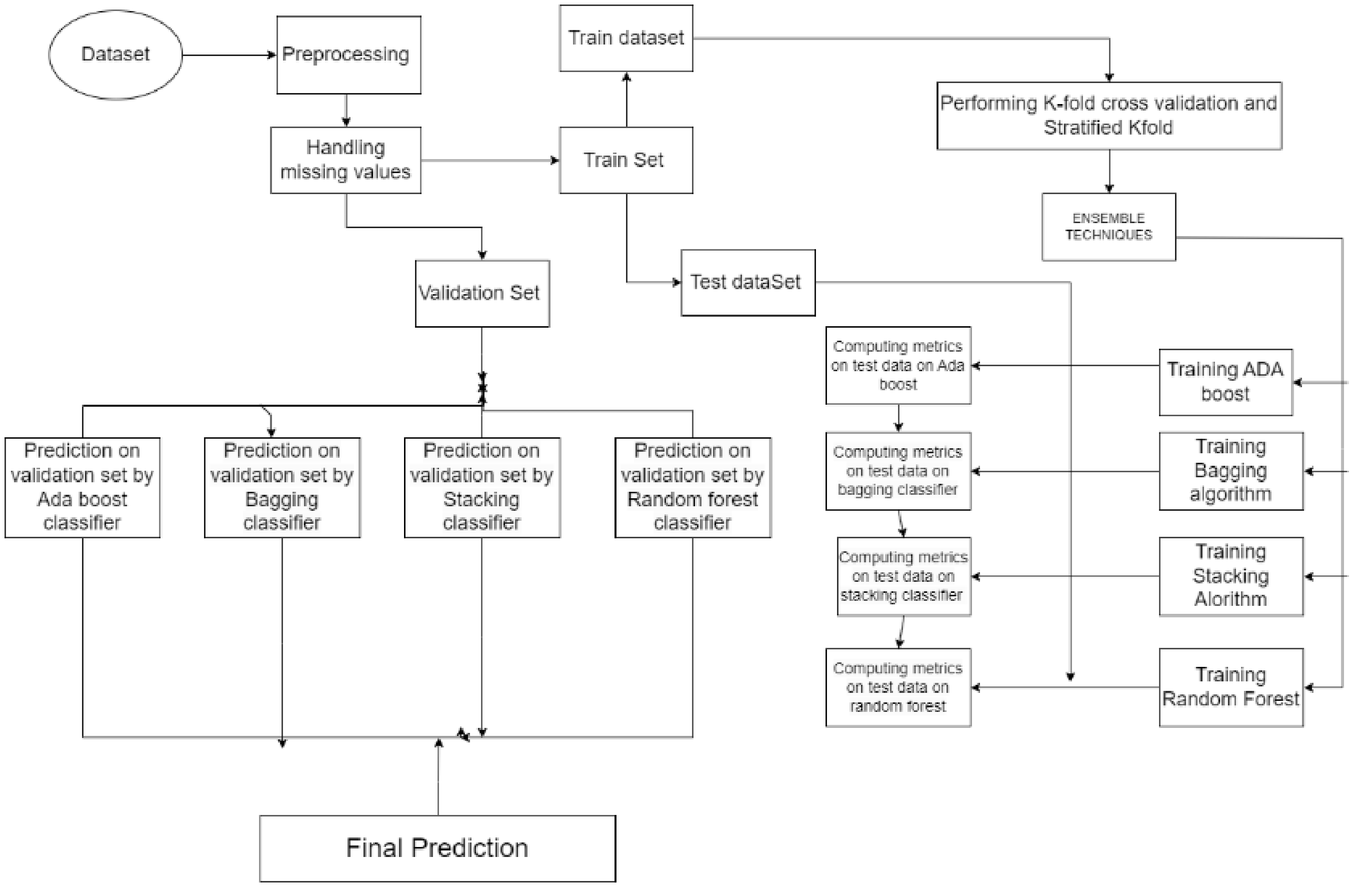

Dataset
Preprocessing
Handling missing values
Train dataset
Train Set
Validation Set
Test dataSet
Performing K-fold cross validation and Stratified Kfold
ENSEMBLE TECHNIQUES
Computing metrics on test data on Ada boost
Computing metrics on test data on bagging classifier
Computing metrics on test data on stacking classifier
Computing metrics on test data on random forest
Training ADA boost
Training Bagging algorithm
Training Stacking Alorithm
Training Random Forest
Prediction on validation set by Ada boost classifier
Prediction on validation set by Bagging classifier
Prediction on validation set by Stacking classifier
Prediction on validation set by Random forest classifier
Final Prediction

Figure 1, (Architecture Diagram)

a) 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.

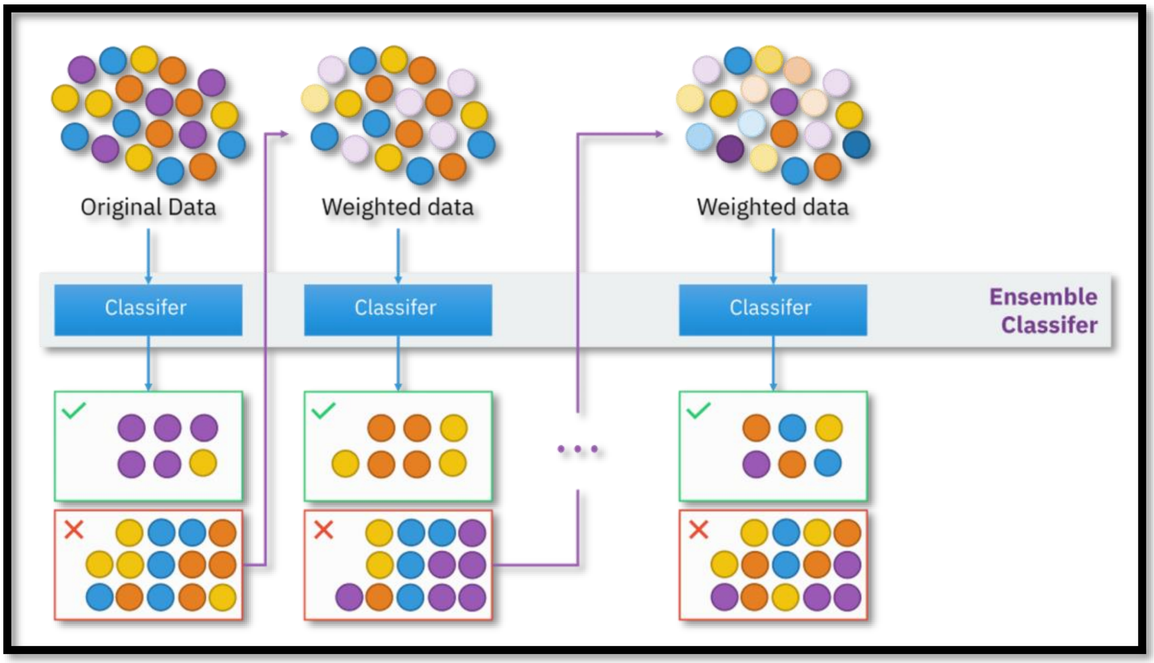


Figure 2, (Ada Boost Working Diagram)

b) **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.

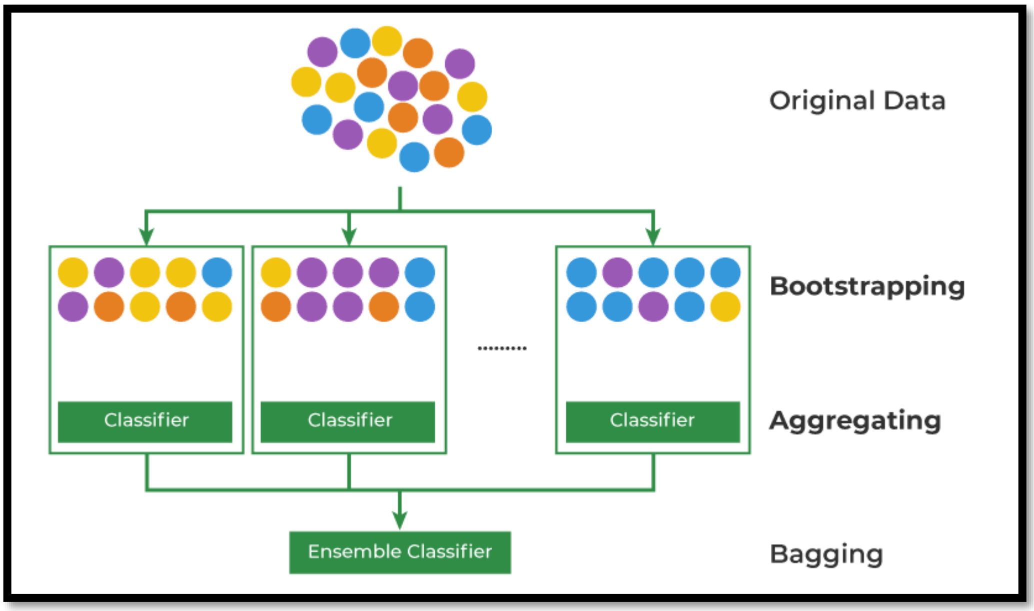


Figure 2, (Bagging Ensemble Working Diagram)

c) **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.

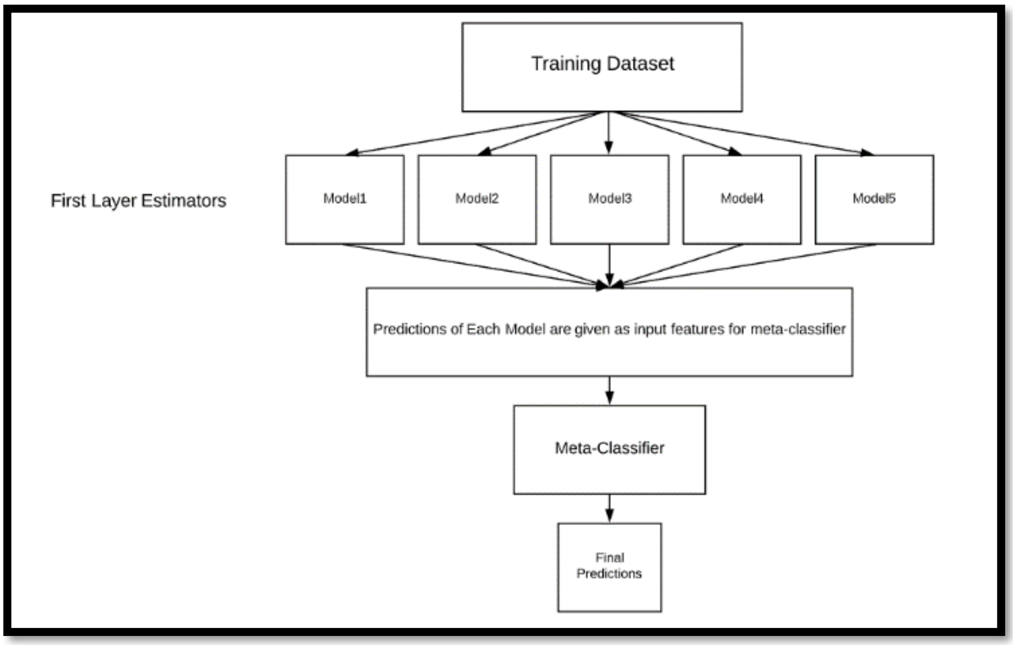


Figure 3, (Stacking Ensemble Working Diagram)

d) **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.

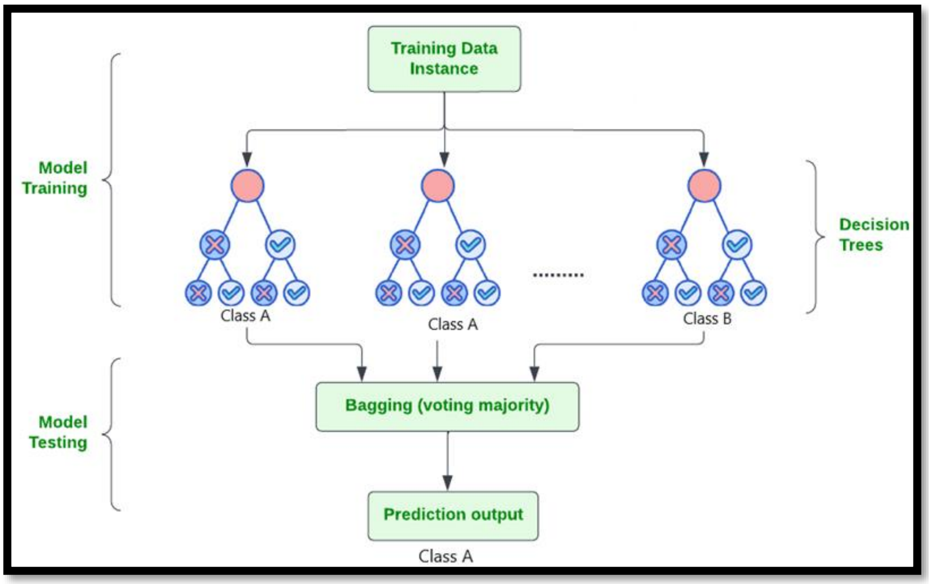
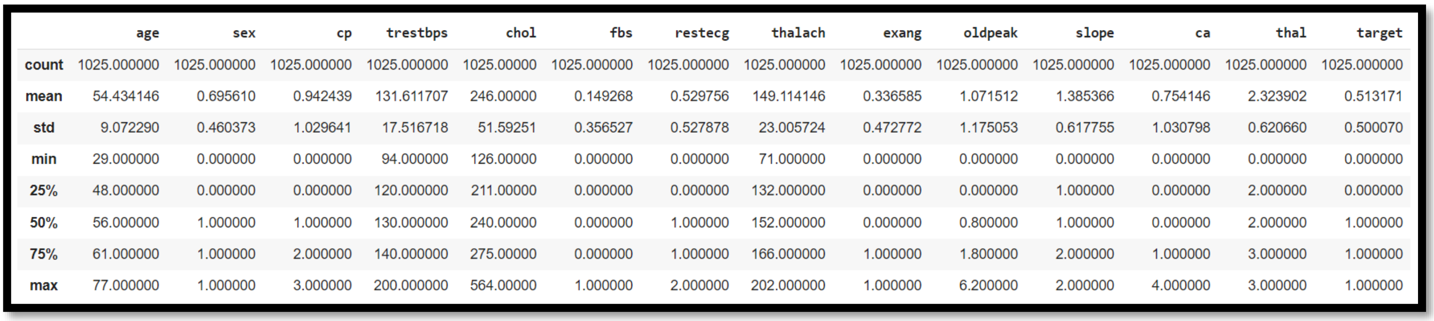


Figure 3, (Random Forest Working Diagram)

RESULTS:

Exploratory Data Analysis (EDA):



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	1025.000000	1025.000000	1025.000000	1025.000000	1025.00000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000
mean	54.434146	0.695610	0.942439	131.611707	246.00000	0.149268	0.529756	149.114146	0.336585	1.071512	1.385366	0.754146	2.323902	0.513171
std	9.072290	0.460373	1.029641	17.516718	51.59251	0.356527	0.527878	23.005724	0.472772	1.175053	0.617755	1.030798	0.620660	0.500070
min	29.000000	0.000000	0.000000	94.000000	126.00000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	48.000000	0.000000	0.000000	120.000000	211.00000	0.000000	0.000000	132.000000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	56.000000	1.000000	1.000000	130.000000	240.00000	0.000000	1.000000	152.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	275.00000	0.000000	1.000000	166.000000	1.000000	1.800000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.00000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

Figure 6, (Data Exploration)

Explanation: The summary statistics in Figure 6 provide a quick overview of the dataset. It includes the count, mean, standard deviation, minimum, maximum, and percentiles of each column. These statistics are essential for understanding the central tendencies, variability, and distribution of the data, which are crucial steps in the exploratory data analysis process.

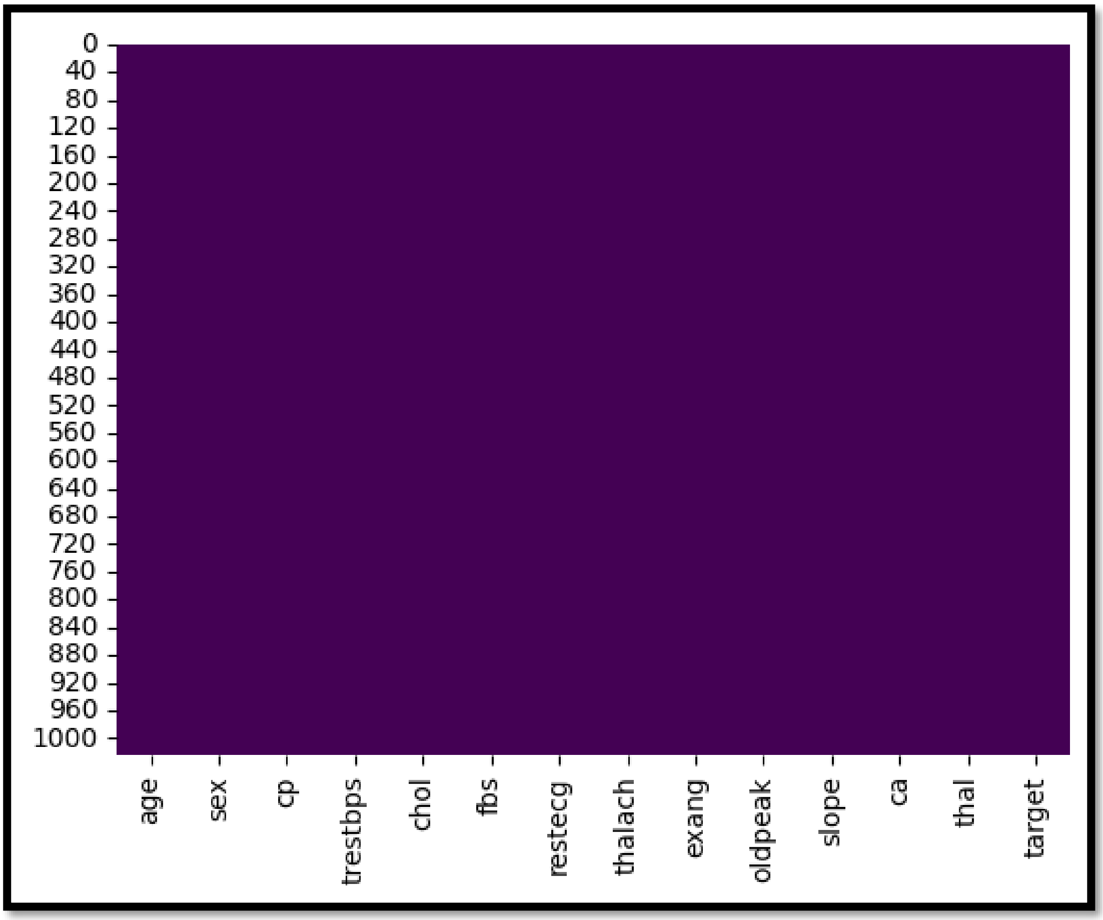


Figure 7, (Visualizing the Missing Values With the help of heatmap)

Explanation: In Figure 7, the heatmap visualization effectively showcases the absence of missing values in the dataset. The heatmap uses colour gradients to represent the presence or absence of data across different columns and rows. Here, the heatmap exhibits a uniform colour pattern, indicating that there are no missing values present. This visualization is crucial in data preprocessing and quality assessment, as it helps ensure that the dataset is complete and ready for further analysis without any data gaps or inconsistencies.

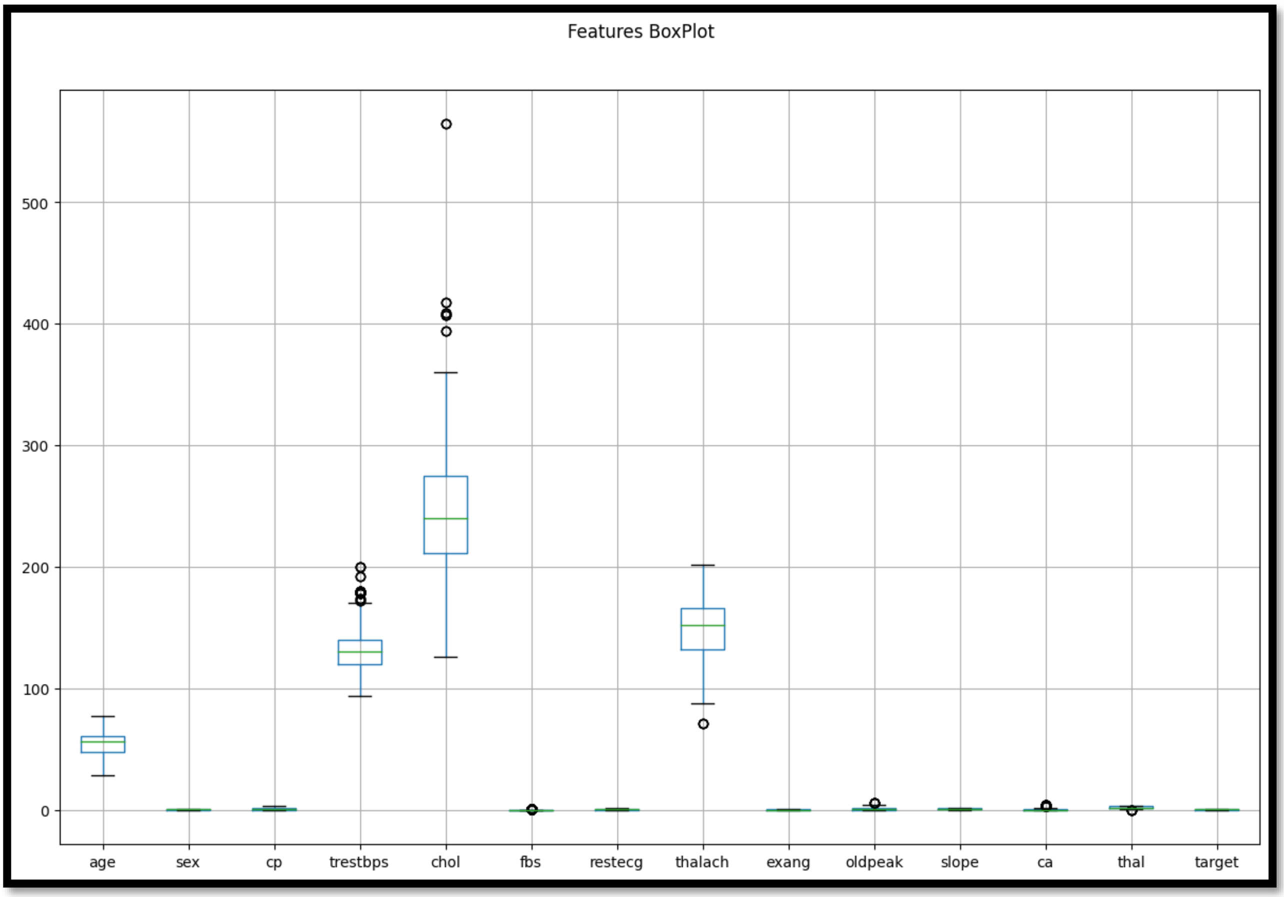


Figure 8, (Boxplot for all features)

Explanation:The boxplot in Figure 8 illustrates the distribution of data for all 14 features in the dataset, including trestbps, chol, fbs, thalach, oldpeak, ca, and thal among others. The presence of outliers in some of these features, such as trestbps, chol, fbs, thalach, oldpeak, ca, and thal, is noticeable. Outliers are data points that significantly differ from the rest of the dataset and can impact the overall analysis. Identifying and handling outliers appropriately is crucial to ensure the reliability and accuracy of statistical analyses and machine learning models built on this data.

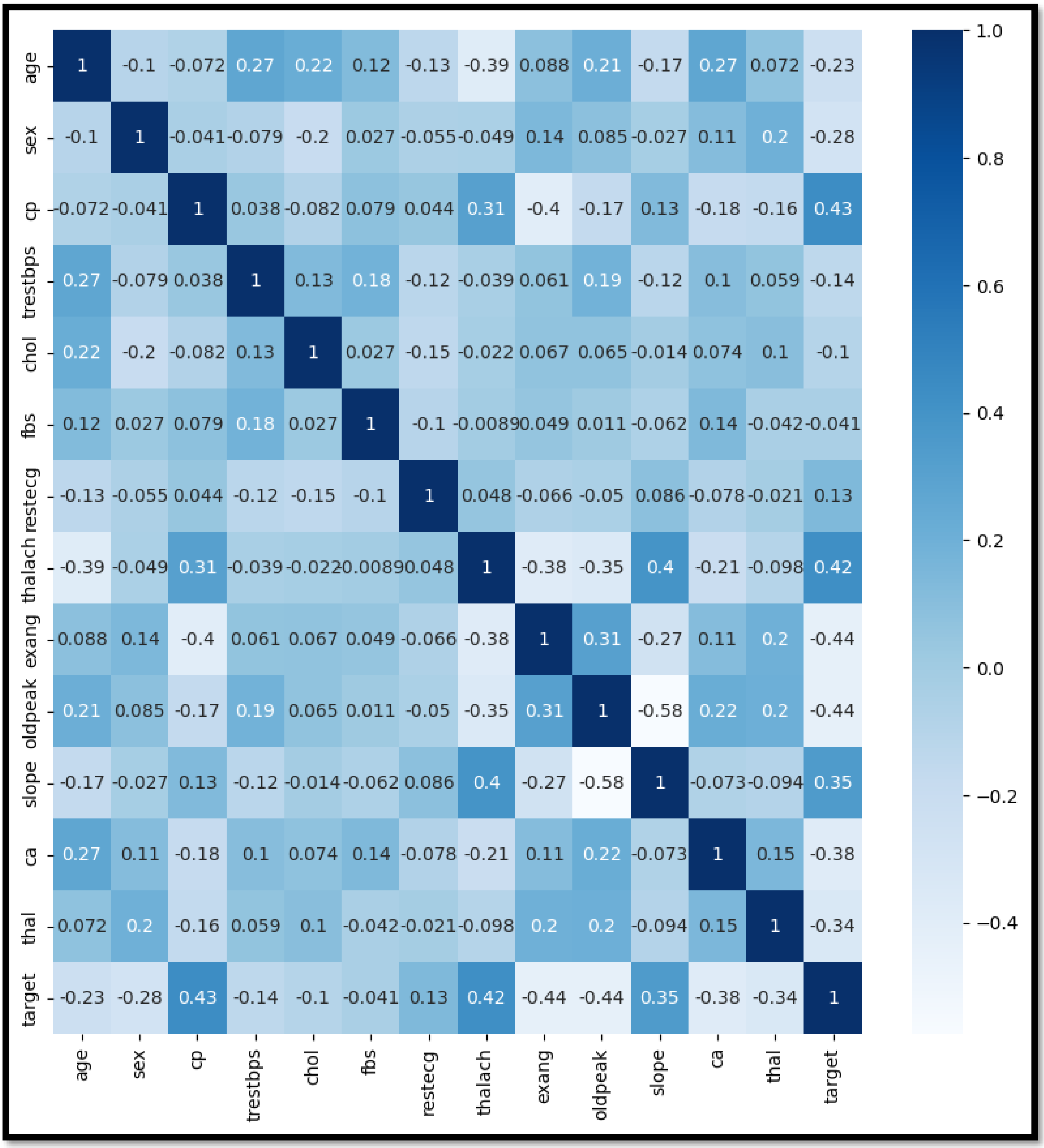


Figure 9, (Correlation between different variables)

Explanation:Figure 9 displays the correlation between various variables and the target column. The positively correlated variables, such as the slope of the peak exercise ST segment, maximum heart rate achieved, resting electrocardiographic results, and chest pain type, tend to increase or decrease together with the target. Conversely, the negatively correlated variables, including age, sex, resting blood pressure, serum cholesterol, fasting blood sugar, exercise-induced angina, ST depression induced by exercise, number of major vessels, and thal, show an inverse relationship with the target column, where one tends to increase as the other decreases, and vice versa. Understanding these correlations is crucial for identifying significant predictors and relationships within the dataset.

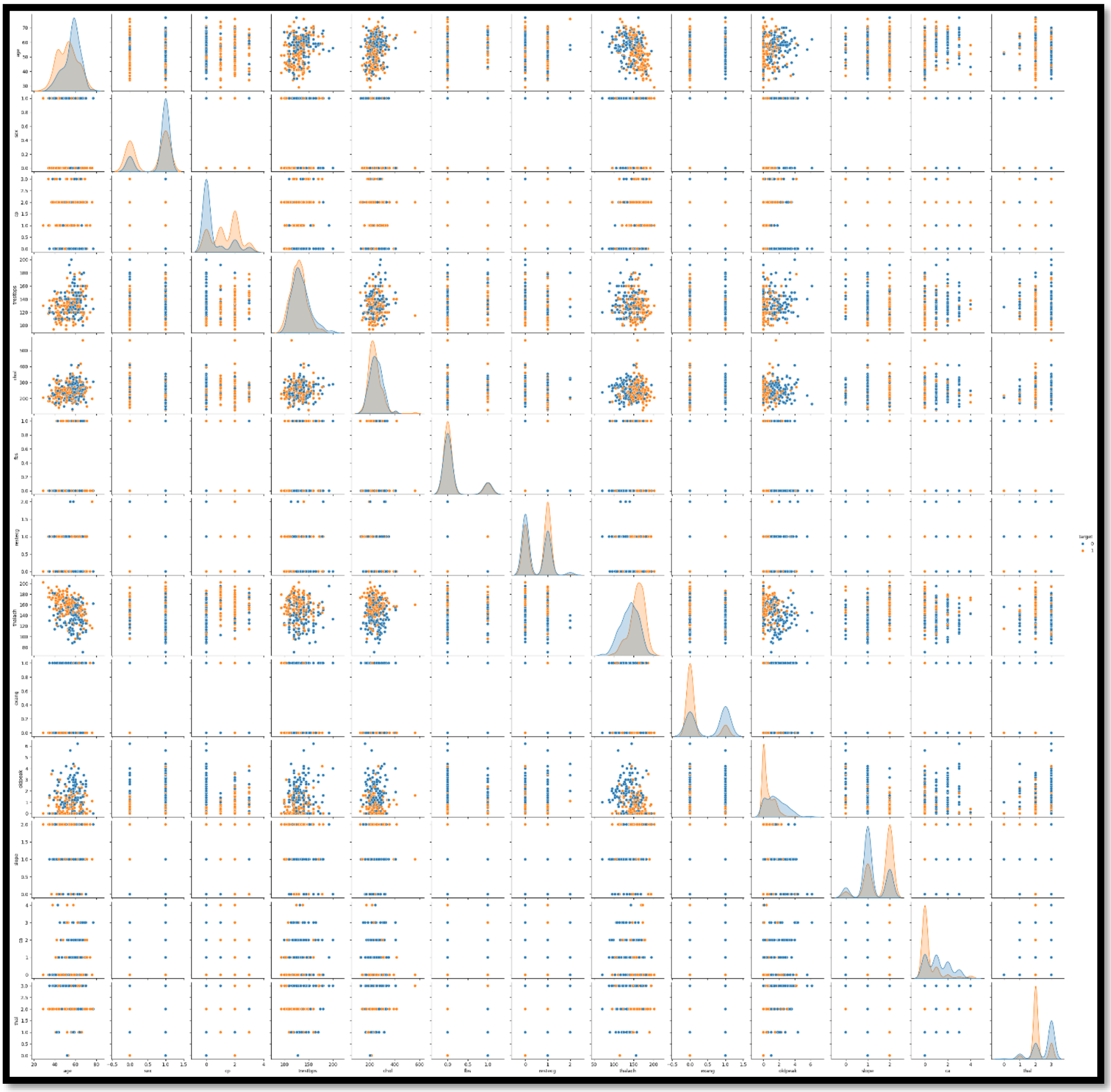
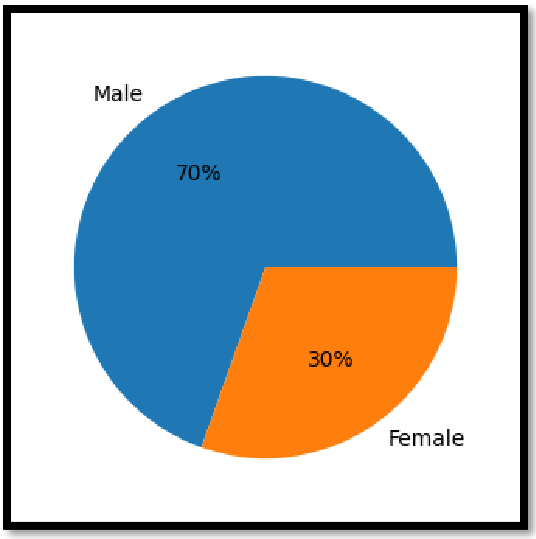


Figure 10,(Pairplot for complete dataframe)

Explanation:This visualization helps in identifying patterns, correlations, and potential insights into the data's structure and characteristics.



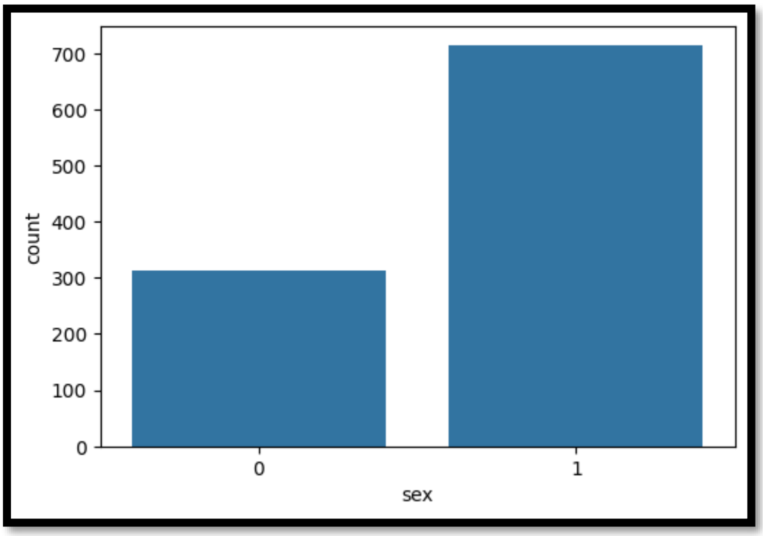


Figure 11 , (Pie-Chart of Sex Column)

Figure 12 , (Count Plot for Sex Column)

Explanation:

The pie chart in *Figure 11* visually represents the distribution of sexes in the dataset of 303 total entries. It reveals that females account for approximately 30% of the dataset, providing a clear snapshot of the gender distribution.

Figure 12 displays a count plot representing the frequency of each sex category in the dataset. It further emphasizes the distribution of males and females, highlighting the relative proportion of each gender within the dataset.

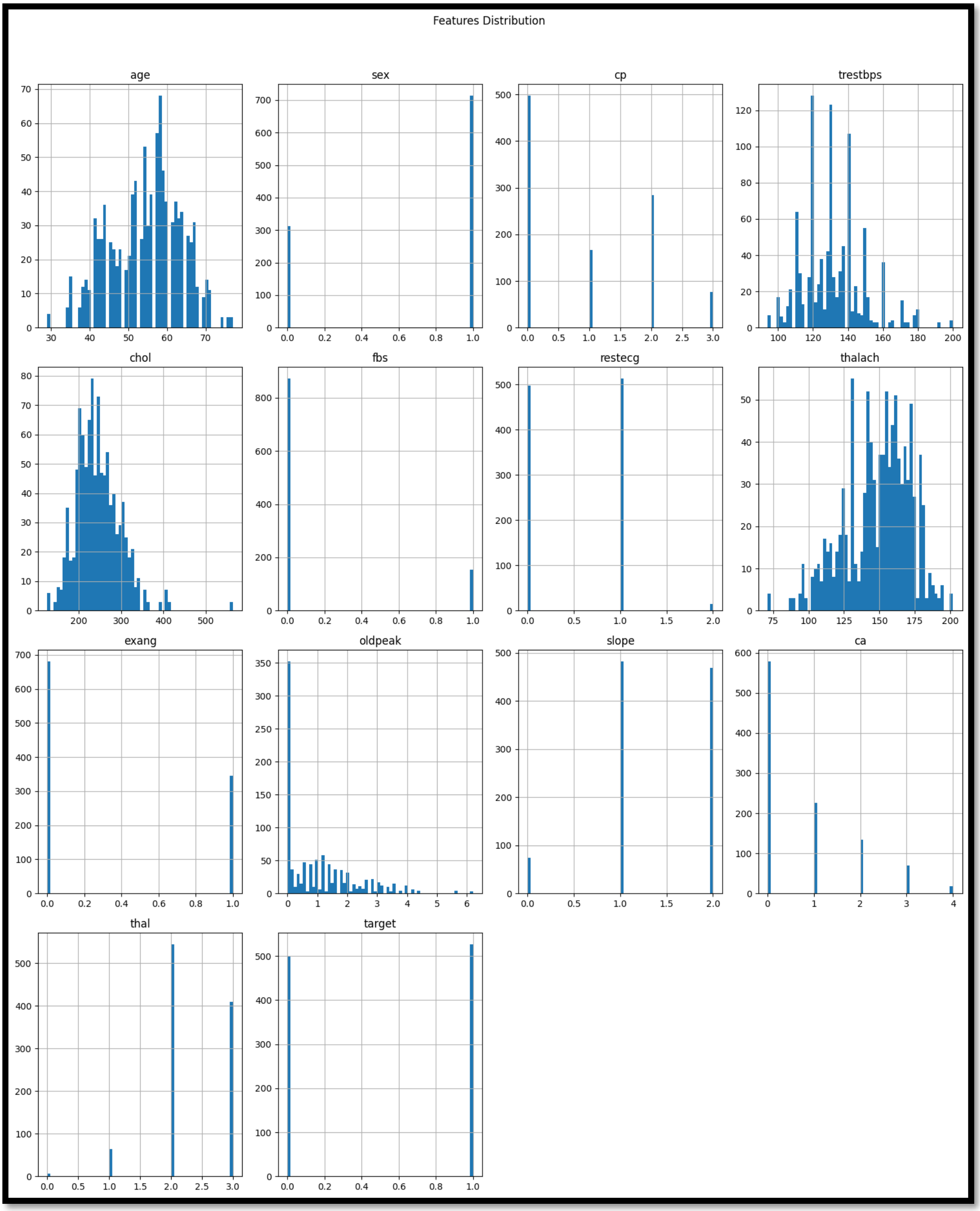


Figure 13, (Features distribution histogram plot)

Explanation: The histogram plot provides a comprehensive view of the distribution of features across the dataset. In this visualization, each histogram represents the distribution of a specific feature, including trestbps, chol, fbs, thalach, oldpeak, ca, thal, age, sex, resting blood pressure, serum cholesterol, fasting blood sugar, and number of major vessels. Analyzing these histograms collectively offers insights into the spread and concentration of data points for each feature, aiding in identifying potential patterns, outliers, and understanding the overall data structure. This consolidated view of feature distributions is valuable for exploratory data analysis, statistical analysis, and informing modeling decisions in various data-driven tasks.

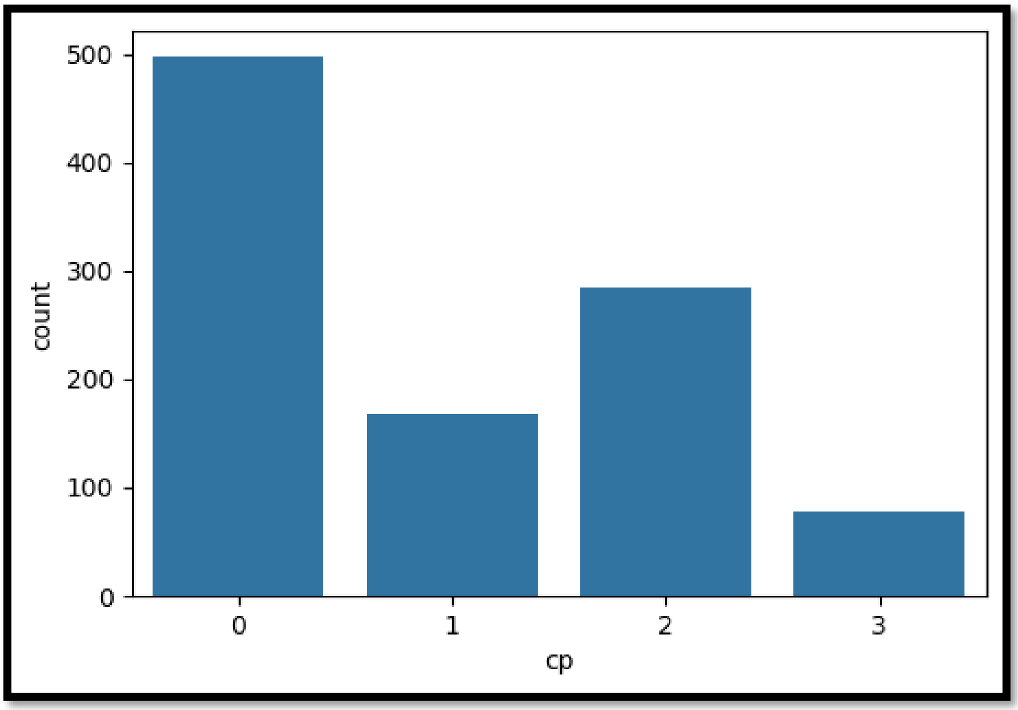


Figure 14,(Count Plot for Chain Pain type column)

Explanation: The Count Plot for Chest Pain type column (Figure 14) illustrates the distribution of chest pain types in descending order based on frequency. Chest pain type 0 appears most frequently, followed by type 2, type 1, and finally type 3, indicating the prevalence of different chest pain descriptions among the dataset. This visualization provides valuable insights into the relative occurrences of each chest pain type, aiding in understanding the distribution patterns within the data.

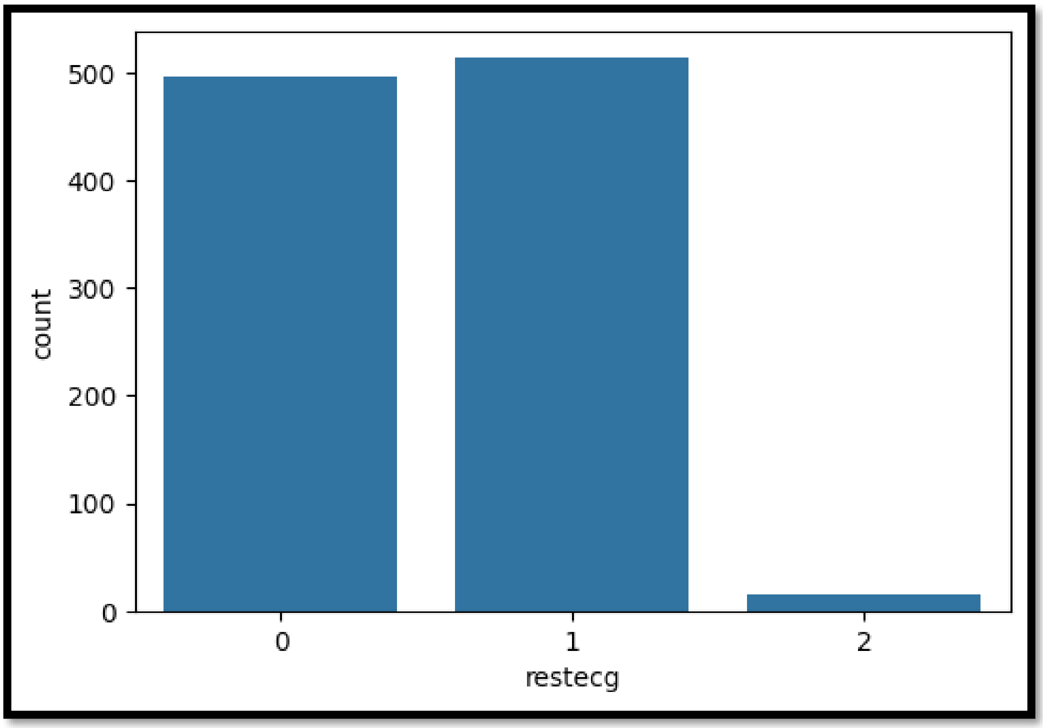


Figure 15, (Count Plot for electrocardiographic results)

Explanation: The Count Plot for electrocardiographic (ECG) results (Figure 15) displays the distribution of ECG outcomes in descending order of occurrence frequency. ECG result 1 is the most prevalent, followed by result 0 and then result 2, showcasing the distribution of different ECG findings within the dataset. This visualization offers a clear representation of the relative frequencies of each ECG result, providing valuable insights into the prevalence of specific ECG patterns among the patients studied.

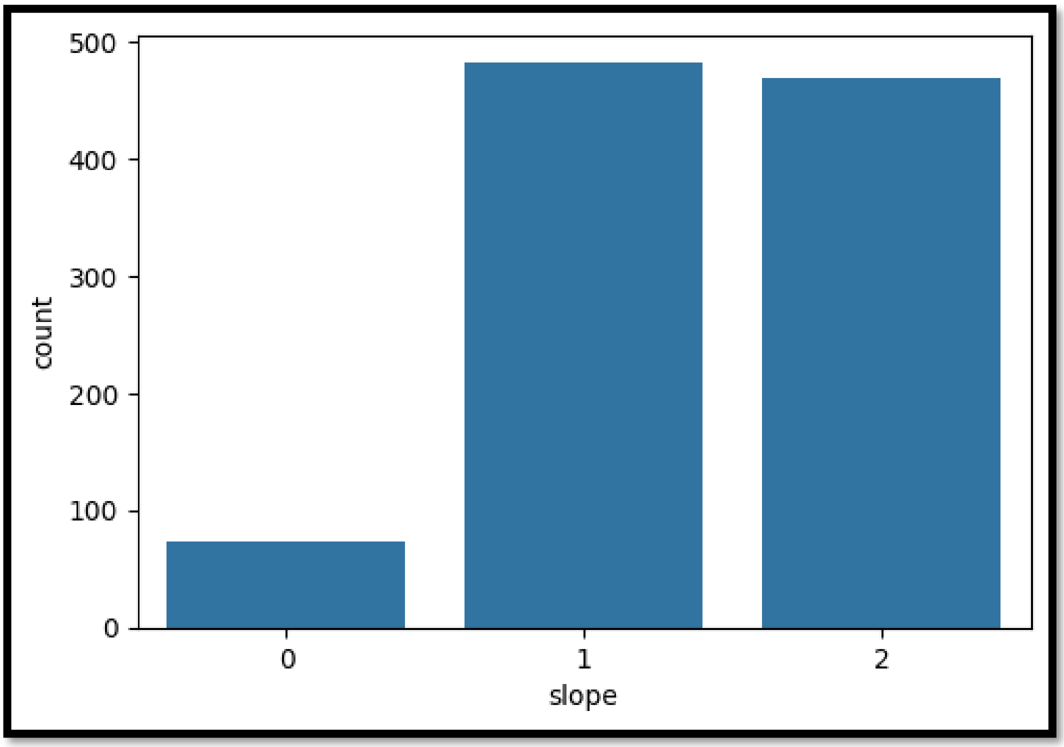


Figure 16, (count plot for st slope column)

Explanation: The Count Plot for the ST slope column (Figure 16) illustrates the distribution of ST slope types in descending order based on their occurrence frequency. ST slope type 1 is the most commonly observed, followed by type 2, and finally type 0, providing a straightforward depiction of the prevalence of different ST slope patterns in the dataset. This visualization helps in understanding the relative frequencies of each ST slope type, offering valuable insights into the distribution of ST segment characteristics among the patients studied.

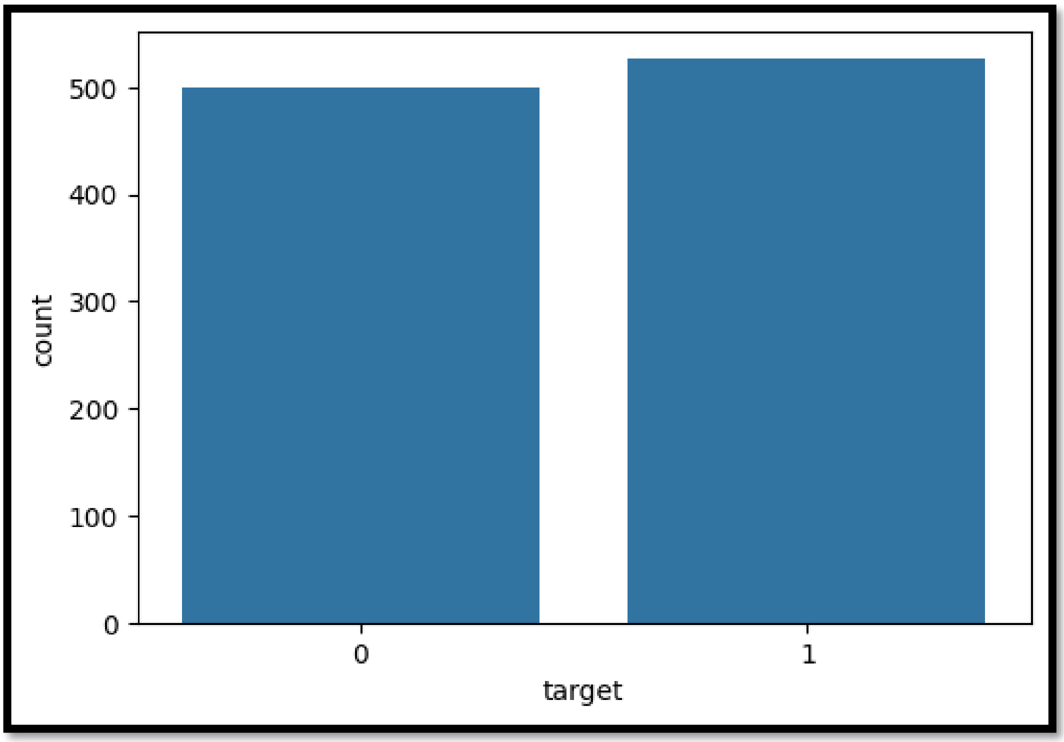


Figure 17, (count plot for target column)

Explanation: The Count Plot for the target column (Figure 17) reveals an almost equal distribution between the values 0 and 1, with a slightly higher frequency of occurrences for the value 1 compared to 0. This visualization underscores the balance in the dataset regarding the presence and absence of the target condition, providing a clear overview of the target variable distribution. It indicates that the dataset contains a substantial number of instances for both target classes, which is essential for training and evaluating predictive models effectively.

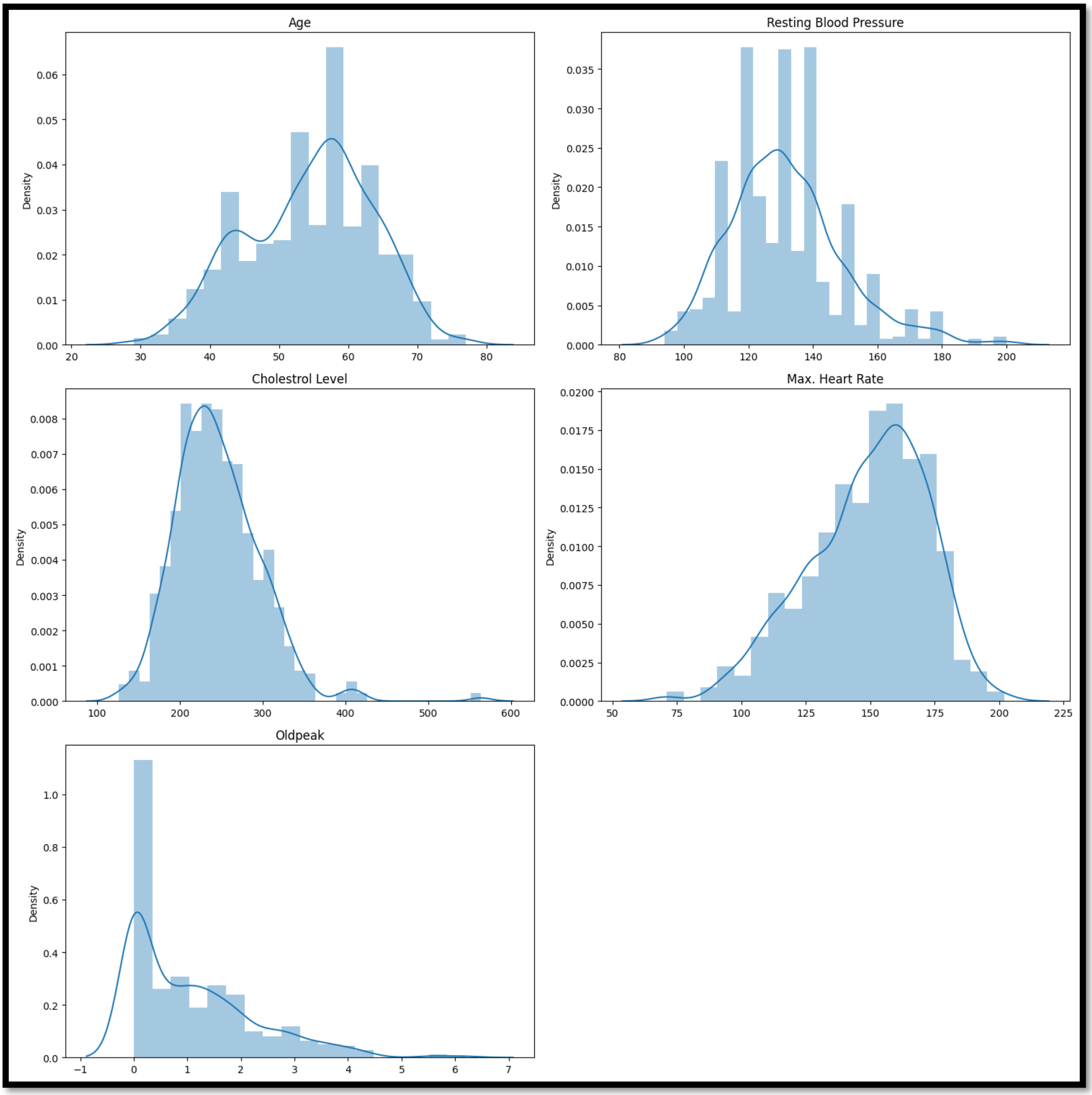


Figure 18, (histogram with kde to see the distribution of features)

Explanation: The histogram with kernel density estimation (KDE) in Figure 18 illustrates the distribution of several key features in the dataset. These features include age, resting blood sugar, cholesterol level, maximum heart rate, and ST depression induced by exercise relative to rest (oldpeak). By visualizing these distributions, we can gain insights into the spread and concentration of values within each feature, which is crucial for understanding the data's characteristics and potential patterns.

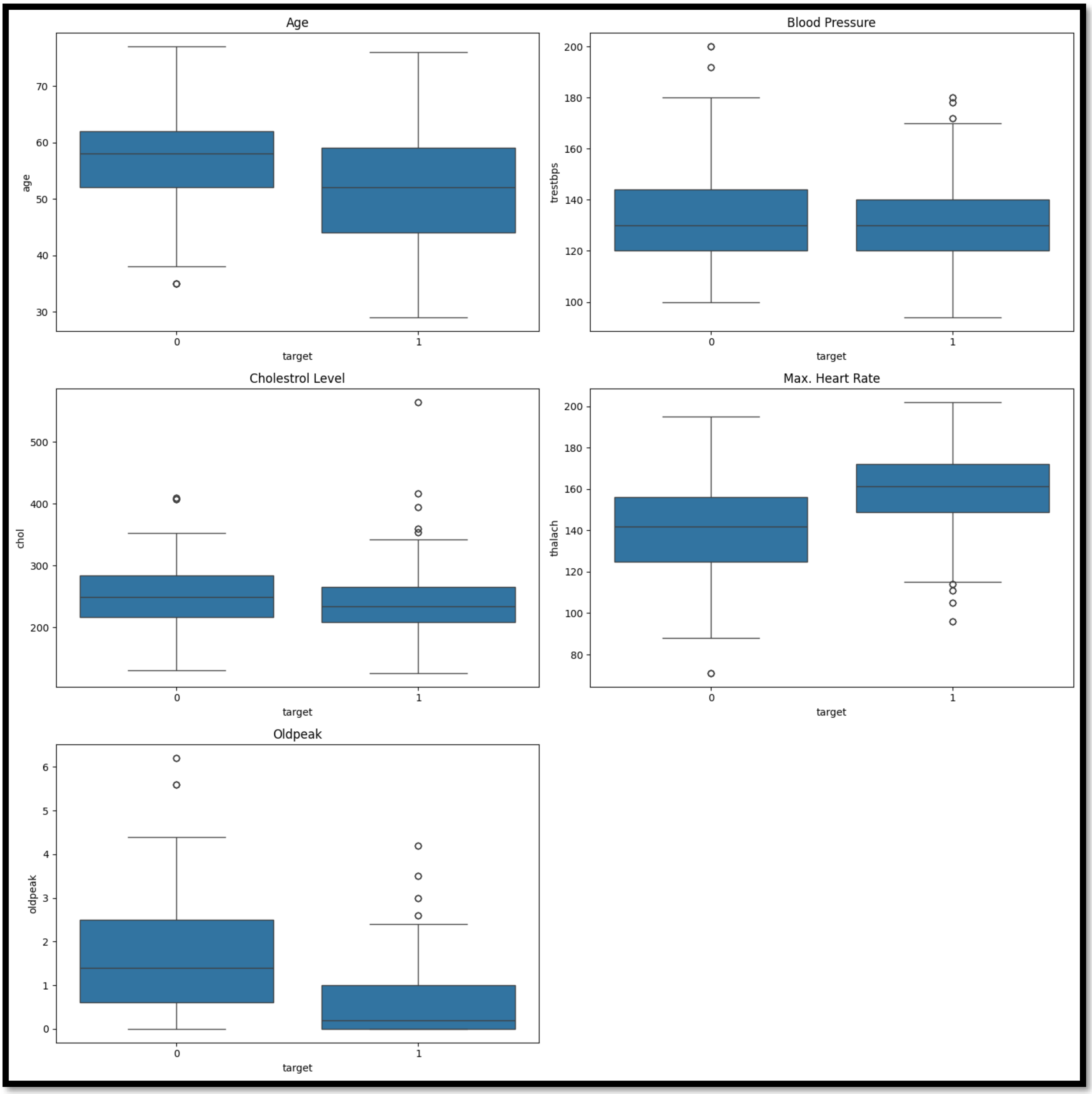


Figure 19, (Boxplots to see the distribution)

Explanation: Figure 19 presents boxplots depicting the distribution of the following features concerning the target column, which has values of 0 and 1. The features included in the boxplots are age, blood pressure, cholesterol level, maximum heart rate, and ST depression induced by exercise relative to rest (oldpeak). Additionally, the boxplots highlight the presence of outliers within these features, with values ranging from 2 to 4 standard deviations from the mean. These outliers provide valuable insights into potential anomalies or extreme values within the dataset, which may warrant further investigation or preprocessing steps during analysis.

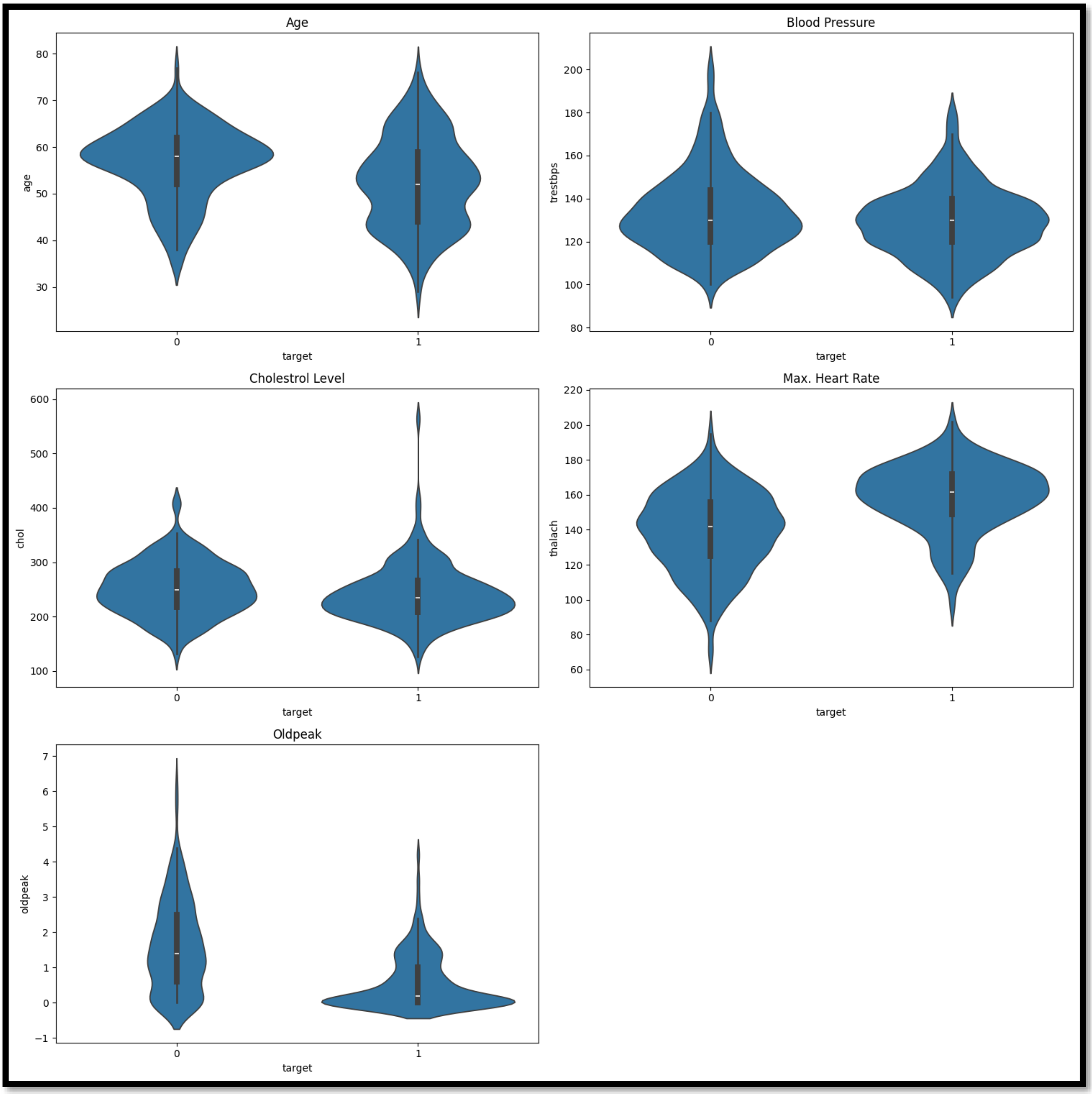


Figure 20, (Violinplots for features ('Heart Disease Result' vs 'All numeric features')

<u>Explanation:</u> Figure 20 illustrates violin plots comparing the 'Heart Disease Result' (target variable) against all numeric features, including age, blood pressure, cholesterol level, maximum heart rate, and ST depression induced by exercise relative to rest (oldpeak). Violin plots are effective in displaying the distribution and density of data points across different categories, providing insights into the relationship between these features and the occurrence of heart disease. This visualization can help identify patterns and trends in how these numeric features contribute to predicting heart disease outcomes.

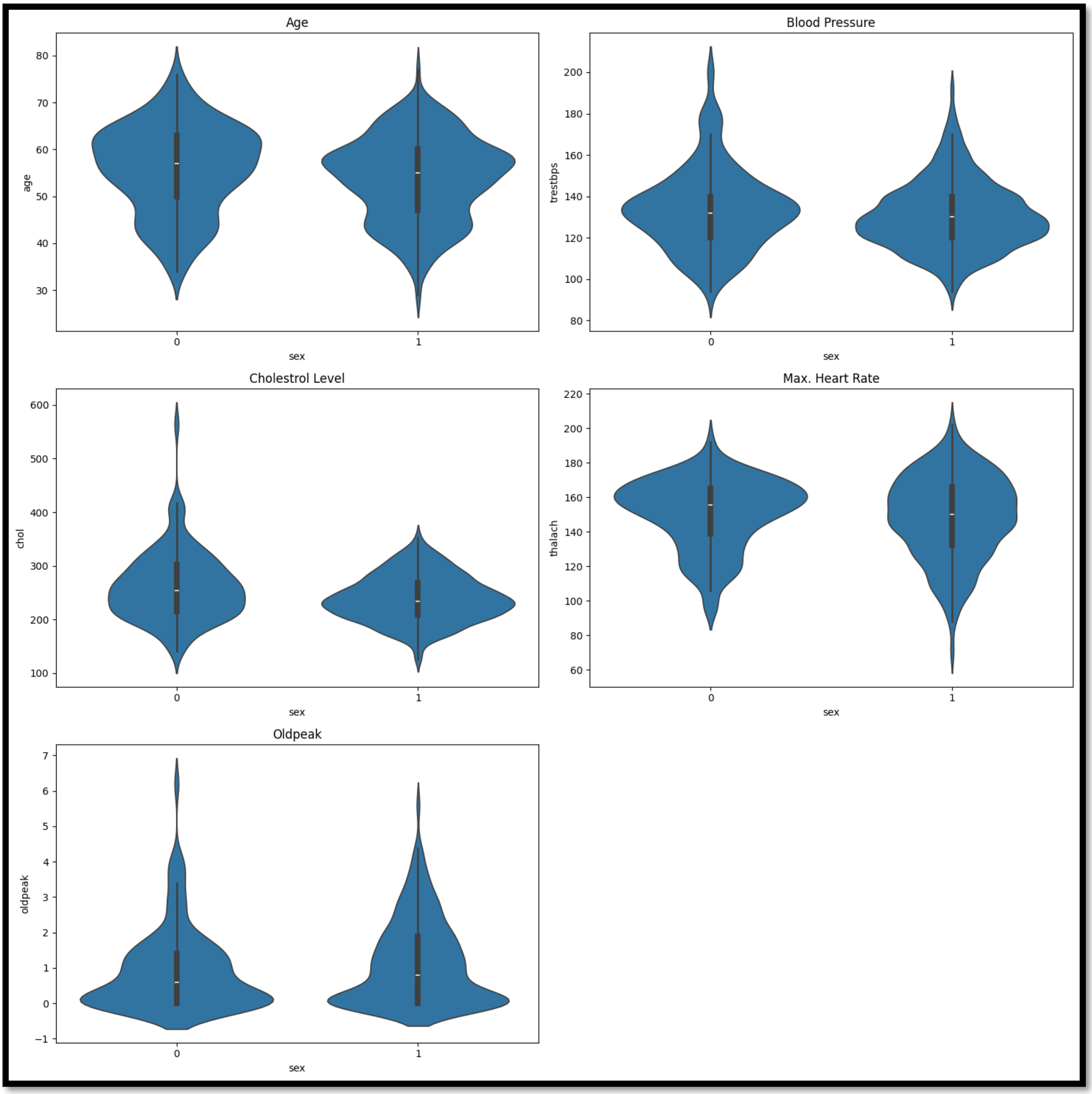


Figure 21, (violinplots for features ('Sex' vs 'All numeric features'))

Explanation: Figure 21 showcases violin plots comparing the 'Sex' attribute against all numeric features, including age, blood pressure, cholesterol level, maximum heart rate, and old peak (ST depression induced by exercise relative to rest). Violin plots are particularly useful for visualizing the distribution of numeric data across different categories, in this case, comparing the distribution of these features between genders (0 for Female and 1 for Male). These plots provide insights into any variations or patterns in the distribution of these attributes based on gender, aiding in understanding potential gender-related differences in the dataset.

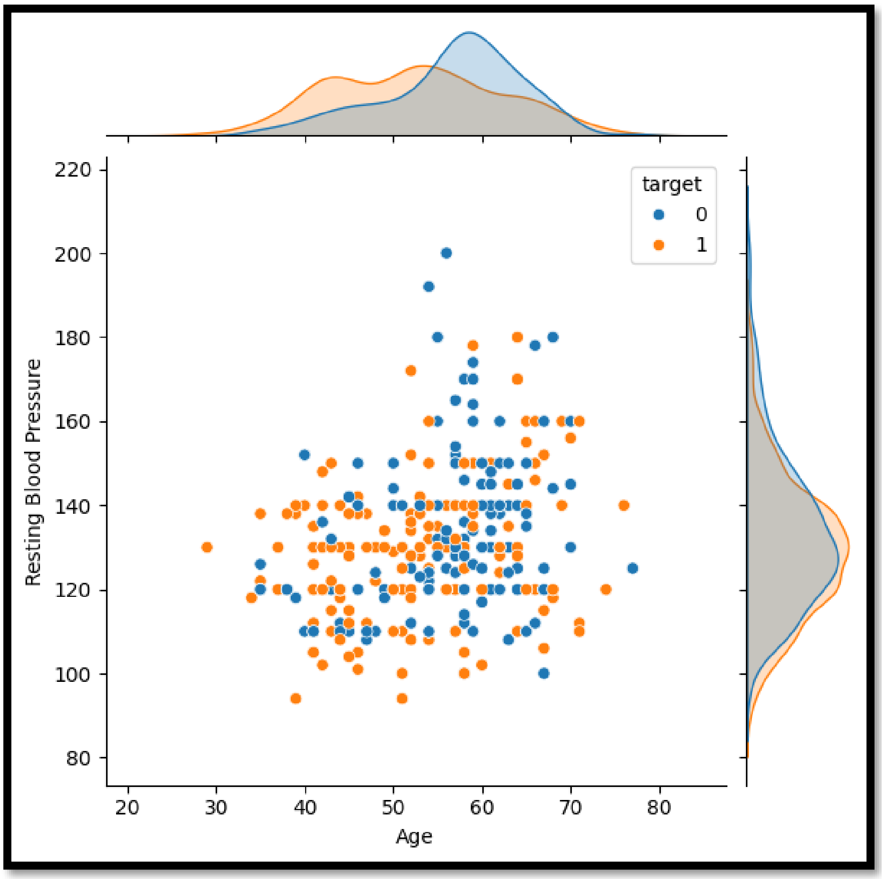


Figure 22, (Jointplot for 'Age' vs 'Max. Heart Rate')

Explanation: The jointplot for 'Age' vs 'Max. Heart Rate' visualizes the relationship between a person's age and their maximum heart rate (thalach) based on the 'target' attribute, which likely denotes the presence or absence of heart disease. The plot uses color hue to distinguish between the target categories, providing a clear depiction of how age and maximum heart rate vary concerning the presence or absence of heart disease. This plot helps in understanding potential correlations or differences in age-heart rate patterns based on heart disease status.

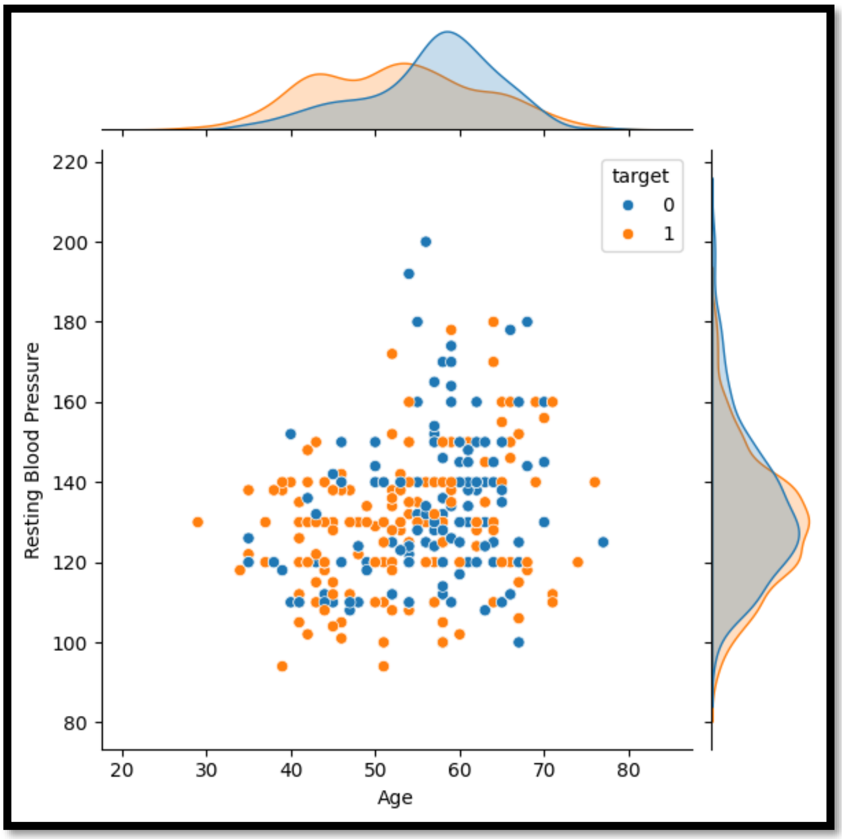


Figure 23 , (Jointplot for 'Age' vs 'Blood Pressure')

Explanation: Figure 23 represents a joint plot comparing the 'Age' and 'Blood Pressure' attributes from the dataset. The hue parameter is set to 'target', which likely indicates that the data points are differentiated based on the target variable (possibly indicating the presence or absence of a certain condition). This joint plot helps visualize the relationship between age and resting blood pressure, with the hue providing additional insight into how this relationship varies concerning the target variable.

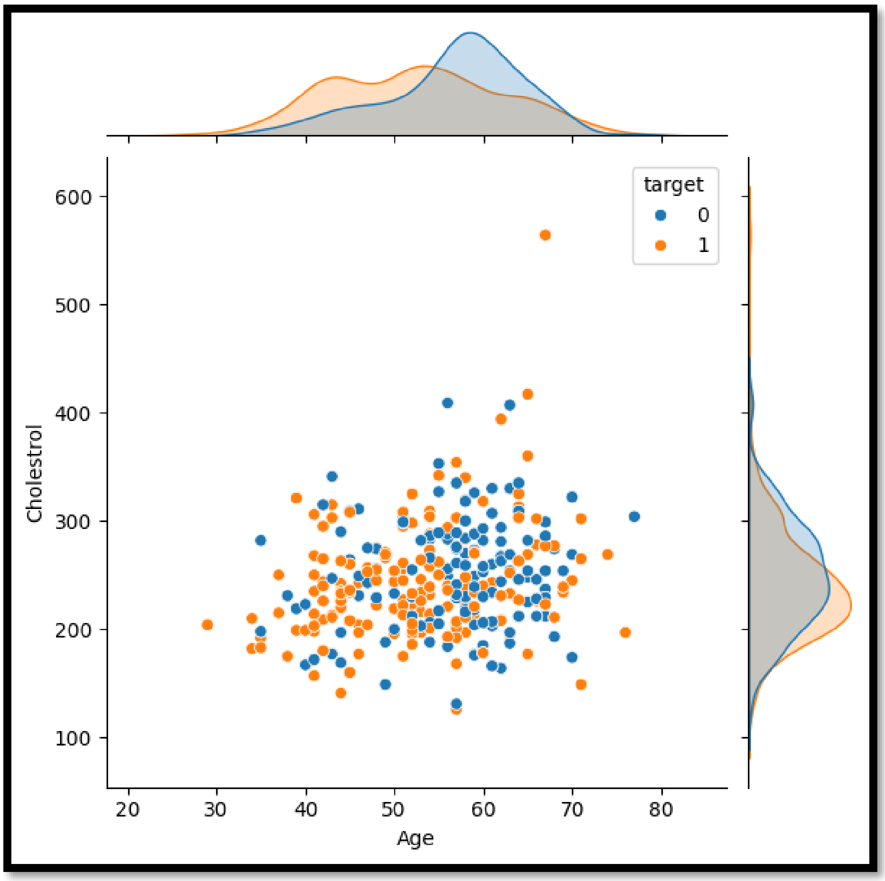


Figure 24, (Jointplot for 'Age' vs 'Cholesterol')

Explanation: The jointplot in Figure 24 illustrates the relationship between age and cholesterol levels ('Cholesterol') in the dataset, with the hue parameter representing the target variable. This plot helps visualize any potential correlation or pattern between a person's age and their cholesterol level, further segmented by the target variable (0 for absence of heart disease and 1 for presence). The scatter points and distribution curves in the jointplot provide insights into how age and cholesterol levels may vary across individuals with and without heart disease, aiding in understanding potential risk factors associated with these variables.

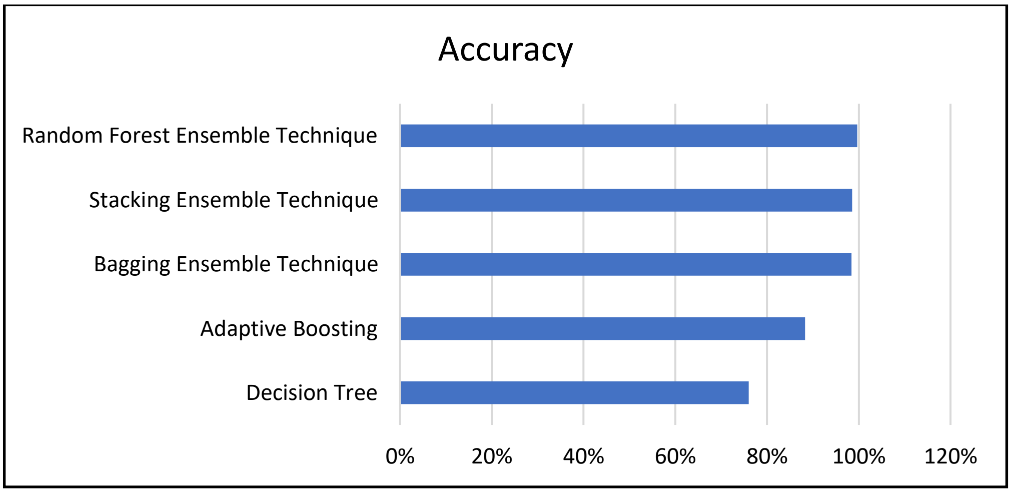


Figure 25, (Accuracy Comparison)

Explanation: Figure 25 provides an accuracy comparison of various ensemble techniques and machine learning models. The results showcase the superior performance of ensemble methods, particularly the Random Forest Ensemble Technique, which achieved an impressive accuracy of 99.70%. This comparison highlights the efficacy of ensemble techniques in improving predictive accuracy for heart disease prediction.

OUTPUT:-

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%

The performance of various ensemble techniques and machine learning models was evaluated for predicting cardiovascular disease. The Decision Tree model achieved an accuracy of 76%, demonstrating its capability as a baseline model. Adaptive Boosting showed improvement with an accuracy of 88.3%, indicating enhanced predictive power. Bagging Ensemble Technique further increased accuracy to 98.44%, showcasing the effectiveness of aggregating predictions from multiple models. Stacking Ensemble Technique demonstrated even higher accuracy at 98.53%, highlighting the benefits of combining diverse models. Finally, the Random Forest Ensemble Technique yielded the highest accuracy of 99.70%, making it the most reliable model for cardiovascular disease prediction among the evaluated techniques.

CONCLUSION:

The research presented in this paper marks a significant advancement in cardiovascular disease (CVD) prediction through the utilization of ensemble techniques and deep learning models. By amalgamating predictions from diverse models such as Decision Trees, Adaptive Boosting, Bagging, Stacking, and Random Forests, our approach has demonstrated remarkable accuracy, with the Random Forest Ensemble Technique achieving an impressive 99.70%. This precision in prediction not only facilitates early detection but also enables proactive prevention strategies and tailored interventions, ultimately reducing the overall cost of CVD treatment. Moreover, the incorporation of big data analytics enhances the system's capability to capture subtle risk factors and improve predictive power.

Looking ahead, the future scope of this research extends towards enhancing users' trust and understanding of the predictive models. Future endeavors should focus on developing interpretability techniques that provide transparent explanations for the model's decisions. This transparency not only boosts users' confidence in the model but also aids in making well-informed health decisions. Additionally, continued collaboration between researchers, data scientists, and healthcare practitioners is essential for driving continuous innovation in cardiovascular healthcare. By leveraging the synergy between advanced predictive modeling techniques and interdisciplinary teamwork, we can revolutionize CVD management, leading to optimized patient outcomes and improved global health.

In summary, the fusion of ensemble learning, deep learning, big data analytics, and collaborative teamwork represents a paradigm shift in cardiovascular healthcare. This shift is characterized by precise predictive models, proactive interventions, and reduced healthcare costs, offering a beacon of hope for millions affected by CVD worldwide.

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/3