A close up of a logo

Description automatically generated

**CAPSTONE (DATA 2206)**

**Data Analytics Report**

TEAM: DATA WIZARDS

PROFESSOR: SAM PLATI

**Submitted by,**

AKSHAY KUTTICKATTU BINU (100729165)

BLESSON ISSAC (100935589)

DILS SIBY (100886198)

IMMANUEL SEBASTIAN (100935313)

TARUN SABU (100913516)

VARSHA REGHU (100930107)

**Submitted on,**

05/04/2024

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| **Contents** | **Page number** |
| 1. Executive summary | 3 |
| 2. Problem description | 4 |
| 3. Data description | 4 |
| 4. Data preparation | 5 |
| 5. Data analysis solution | 12 |
| 6. Conclusions | 15 |
| 7. Appendix | 16 |

**EXECUTIVE SUMMARY**

**Business Problem**:

Heart diseases cause a significant challenge globally, contributing to high mortality rates and imposing substantial financial burdens on healthcare systems. Traditional diagnostic methods often fail to provide timely risk assessments, demanding more efficient predictive models. Our project aims to address this challenge by leveraging machine learning techniques to identify individuals at risk of heart disease based on clinical and demographic data.

**Data**:

We obtained an extensive dataset from a prominent multispecialty hospital in India, comprising clinical and demographic characteristics of 1000 individuals. This dataset serves as a valuable resource for cardiovascular research, containing information on factors such as age, gender, blood pressure, cholesterol levels, and more.

**Key Metric: Recall**

Recall is identified as the key metric for success in our project. High recall ensures that the model effectively identifies individuals at risk of heart disease among all those who genuinely have it, minimizing false negatives and enabling prompt interventions to mitigate disease progression. Therefore, we prioritize maximizing recall to enhance the effectiveness of our predictive models and improve overall patient care.

**Analytics Solution**:

Our solution involves the development of predictive models using sophisticated machine learning algorithms. We preprocess the data to handle missing values, duplicates, and outliers before conducting exploratory data analysis to identify patterns and insights. Through feature engineering and model selection, we aim to create models that reliably predict the presence or absence of heart disease.

**Recommendations**:

* Early Detection: Implement predictive models to enable early detection of heart disease, allowing for timely interventions and improved patient outcomes.
* Integration into Clinical Practice: Integrate predictive models into clinical workflows to streamline diagnostics and facilitate proactive management of heart-related conditions.
* Healthcare Decision-Making: Provide data-driven evidence to healthcare decision-makers for planning services, allocating resources, and implementing preventative measures.
* Public Awareness: Increase public awareness of cardiovascular health through informed lifestyle decisions based on data-driven insights.

**PROBLEM DESCRIPTION**

**Business Goal:**

The primary objective of our project is to address the significant global burden of heart disease by developing predictive models for early detection. Heart disease is a leading cause of mortality worldwide, placing a substantial financial strain on healthcare systems and negatively impacting public health. The aim is to improve patient outcomes and reduce healthcare costs through timely intervention and personalized treatment plans.

**Data Analysis Goal:**

Our data analysis goal is to leverage machine learning techniques to analyze the Cardiovascular Disease Dataset obtained from a prominent multispecialty hospital in India. We aim to develop accurate and reliable predictive models that can identify individuals at risk of heart disease based on their clinical and demographic characteristics. By analyzing the dataset and applying appropriate algorithms, we seek to create reproducible model outcomes that can assist healthcare providers in making informed decisions and implementing proactive measures for heart disease prevention and management.

**Data Description**

**Size and Dimension:**

* The Cardiovascular Disease Dataset consists of data from one thousand individuals.
* The dataset contains twelve distinct features related to cardiovascular health, along with one target variable indicating the presence or absence of heart disease.

**Type of Data:**

* The dataset includes both numeric and categorical variables.
* Numeric variables: age, restingBP, serumcholesterol, maxheartrate, Oldpeak, noofmajorvessels.
* Categorical variables: chestpain, restingrelectro, slope
* Binary categorical variables: gender, fastingbloodsugar, exerciseangia, target

**Output Variable:**

Target: Indicates the presence (1) or absence (0) of heart disease. This variable serves as the target variable for predictive modeling.

**Input Variables:**

* age: Age of the patient in years.
* gender: Gender of the patient (male or female). (0= female, 1 = male)
* chestpain: Reflects the type of chest pain experienced by the patient (0 for typical angina, 1 for atypical angina, 2 for non-anginal pain, 3 for asymptomatic).
* restingBP: Patient's blood pressure measured in millimeters of mercury (mmHg).
* serumcholestrol: Patient's serum cholesterol level measured in milligrams per deciliter (mg/dL).
* fastingbloodsugar: Indicates the patient's fasting blood sugar level (0 for false, 1 for true).
* restingrelectro: Indicates the resting electrocardiogram (ECG) of the patient (0 for normal, 1 for ST-T wave abnormality, 2 for probable or definite left ventricular hypertrophy).
* maxheartrate: Value of the maximum heart rate achieved by the patient.
* exerciseangina: Indicates whether the patient experienced exercise-induced angina (0 for no, 1 for yes).
* oldpeak: ST depression induced by exercise relative to rest.
* slope: The slope of the peak exercise ST segment (1 for upsloping, 2 for flat, 3 for downsloping).
* noofmajorvessels: Indicates the number of major vessels (0-3) colored by fluoroscopy.

**DATA PREPARATION**

**Data Collection:**

The dataset, consisting of clinical and demographic information for 1000 individuals, was sourced from a cardiovascular health database at a renowned multispecialty hospital in India. This dataset forms the foundation for the subsequent analysis.

**Initial Data Exploration:**

A preliminary examination of the dataset was conducted to gain insights into its structure, variable types, and to identify any noticeable data quality issues. This step helps in understanding the dataset's characteristics before delving into further analysis.

**Data Cleaning and Transformation:**

* + **Data Reduction:** The 'patientid' column, which does not contribute to the analysis, was removed from the dataset to simplify the data structure.
  + **Handling Inconsistencies:** Data integrity was ensured by identifying and addressing character mistakes in the dataset. Rows with inconsistent values for the 'slope' feature were identified and removed to maintain data consistency.
  + **Handling Missing Values:** The dataset was checked for missing values in numeric columns. No missing values were detected, ensuring completeness and accuracy in the dataset.
  + **Outlier Removal:** Outliers, if any, were visualized using box plots. However, no outliers were found, indicating that extreme values did not significantly impact the dataset.
  + **Feature Encoding:** Nominal variables such as 'chestpain', 'restingrelectro', and 'slope' were encoded using one-hot encoding. This transformation converts categorical variables into a numerical format suitable for analysis.
  + **Balancing:** Addressed class imbalance using the Synthetic Minority Over-sampling Technique (SMOTE). This technique ensures that both classes (presence and absence of heart disease) are equally represented in the dataset, enhancing the reliability of the analysis.

**EDA (Exploratory Data Analysis) Approach**

* + 1. **Data Cleaning:**

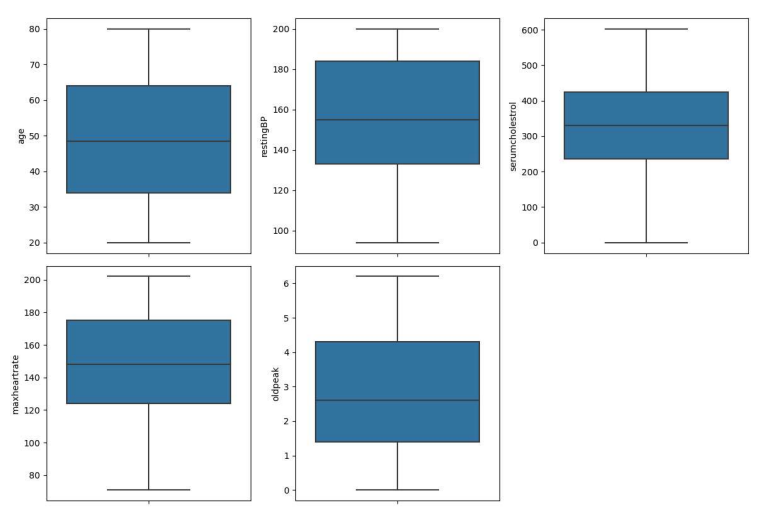
The first step involved cleaning the dataset to handle missing values, duplicates, and inconsistencies.

* + 1. **Univariate Analysis:**

Each variable was individually analyzed to understand its distribution, central tendency, and spread. For this we made use of histograms and boxplots.

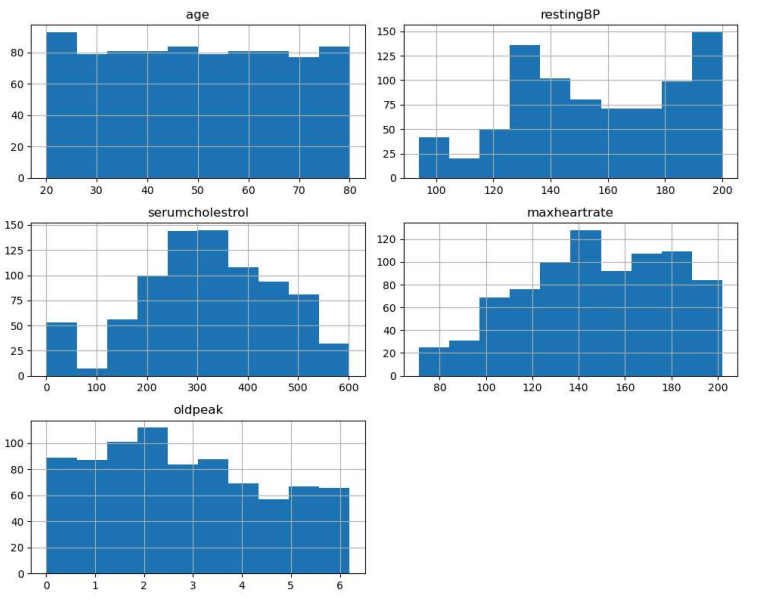
* **Box plot:**

The median, quartiles, and any outliers for each variable are clearly visible when displayed using box plots. The dataset we have appears to be free of outliers. While the classic peak box plot shows a right-skewed distribution, the resting blood pressure and maximum heart rate box plots show left-skewed distributions. It is noteworthy that the box plot for serum cholesterol indicates distribution that is roughly symmetric, pointing in the direction of normalcy.

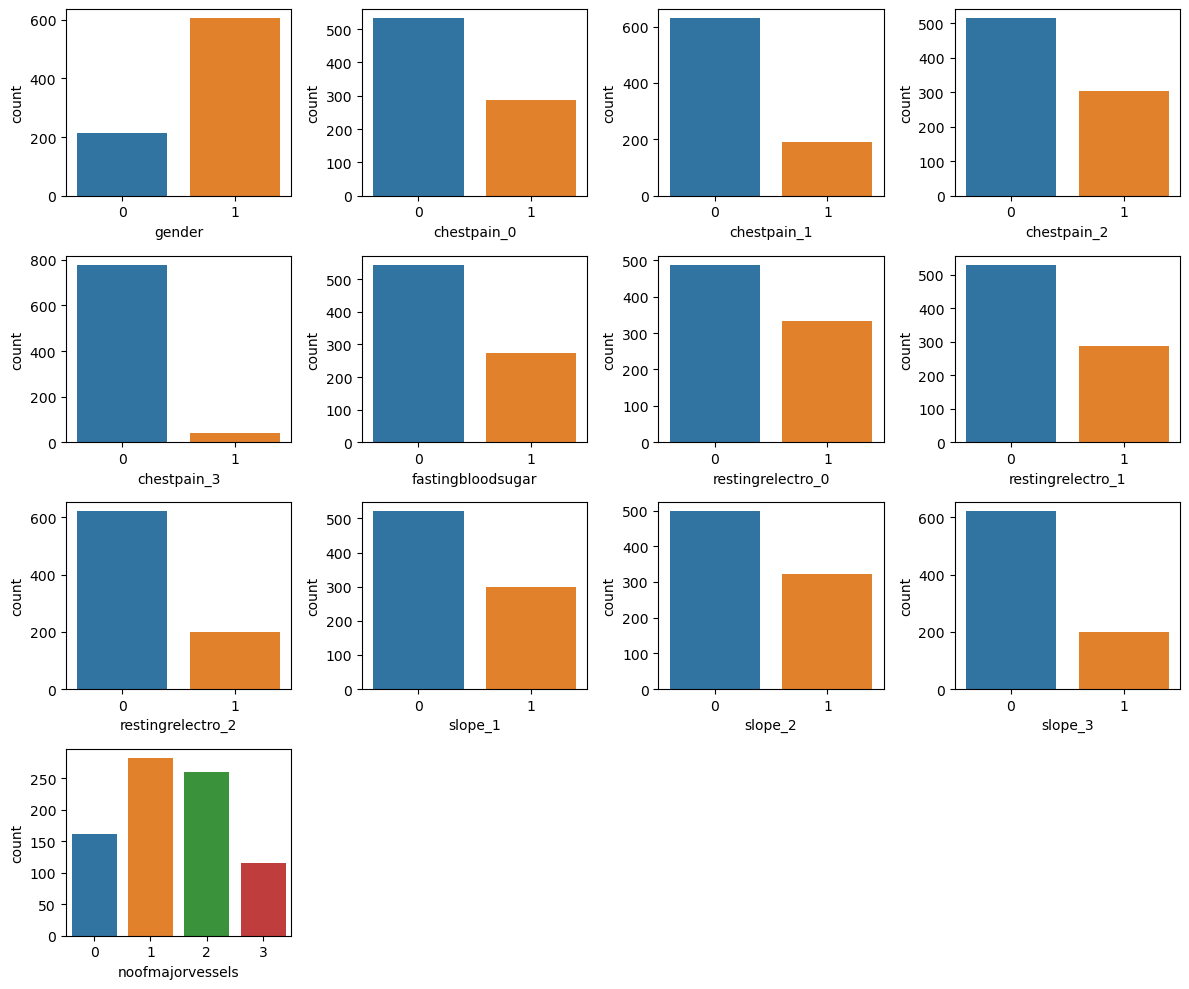


* **Histogram**

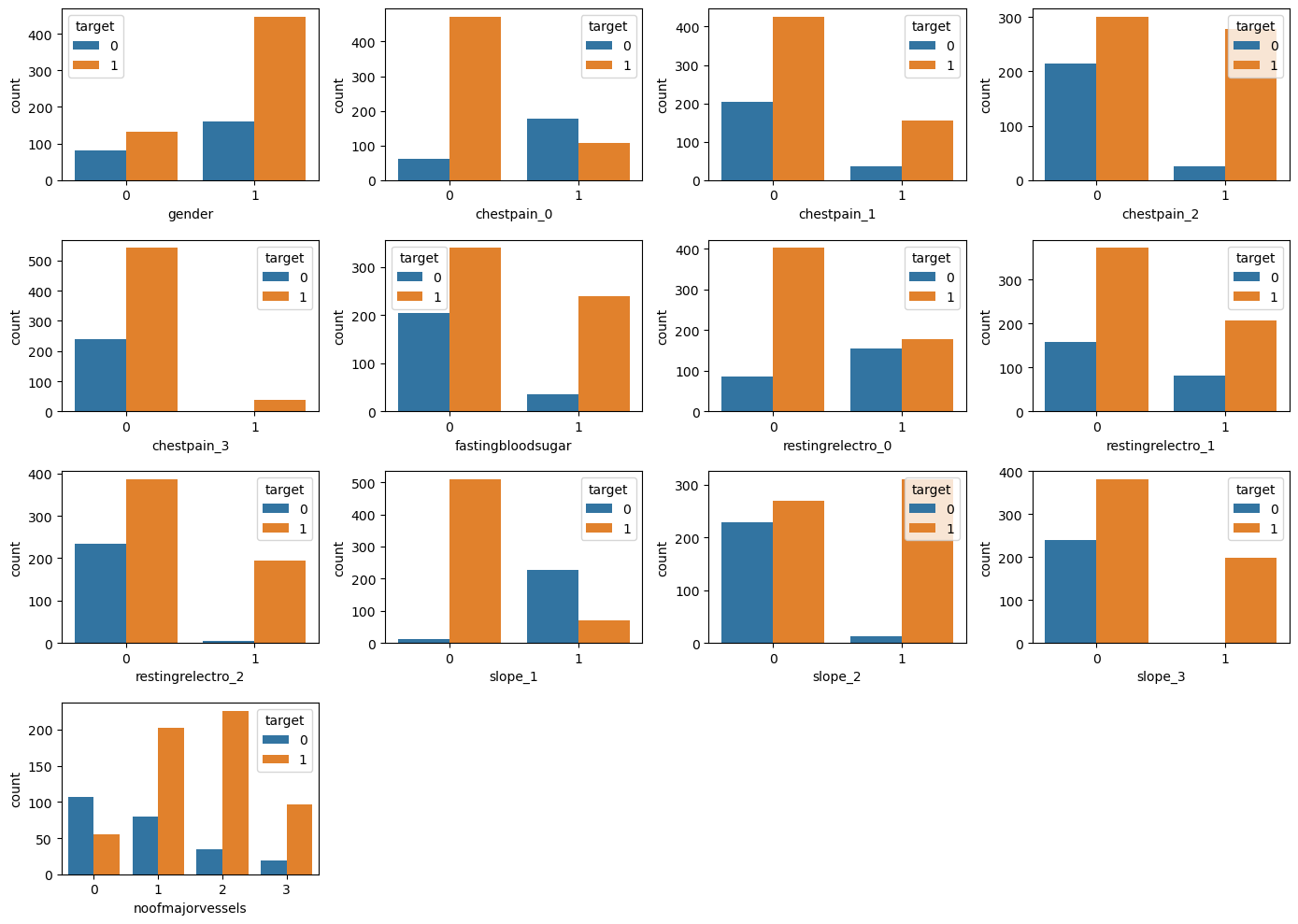
You can see each variable's distribution and spread by looking at its histograms for numerical variables. A bell-shaped curve, for instance, indicates a normal distribution in the histogram of a measure like serum cholesterol. Comprehending the variable distribution is essential for additional examination and modeling since it offers perceptions into the fundamental attributes of the dataset.Also the range of each variable can be determined from the plot. For example it can be noted that patients' ages range from 20 to 80 years.



Likewise, this is the plotted distribution of binary/one-hot encoded data.The following list includes some conclusions drawn from the chart.With 0 representing female and 1 representing male, the bar chart illustrates the gender distribution within the dataset. Clearly, there are more men than women. It's noteworthy that individuals most frequently experience type 2 chest discomfort, whereas type 3 seems to be the least common in this regard.

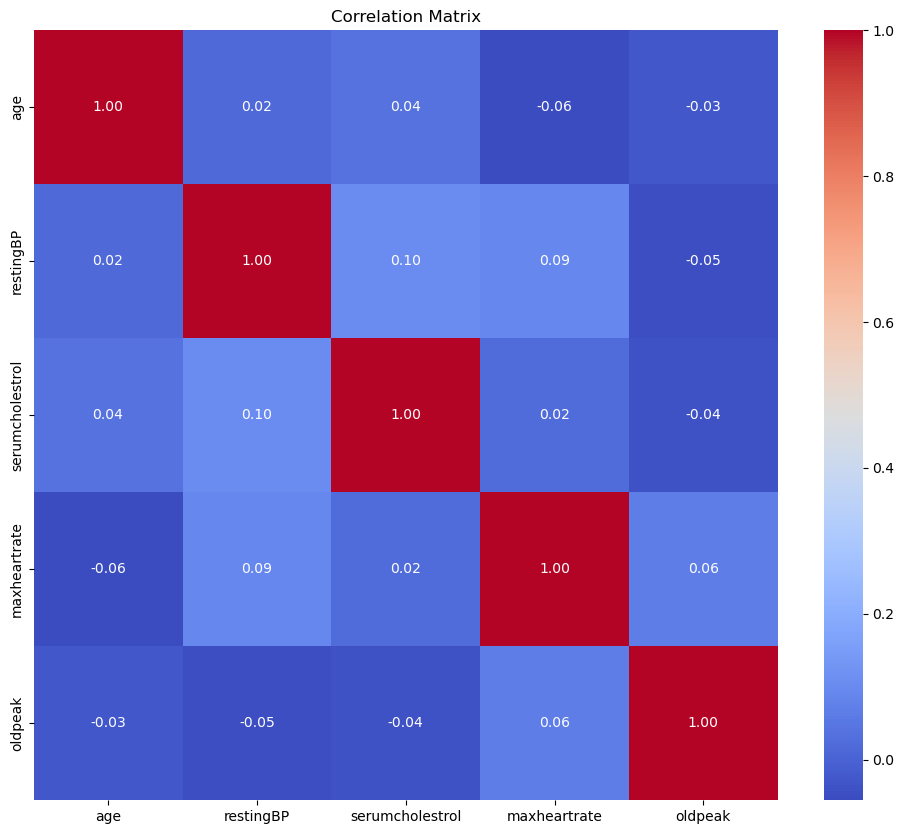


Distribution of categorical variables against target variable is plotted as below.It is observed that heart disease is observed more in males. It is observed that everyone with chest pain 3 has heart disease



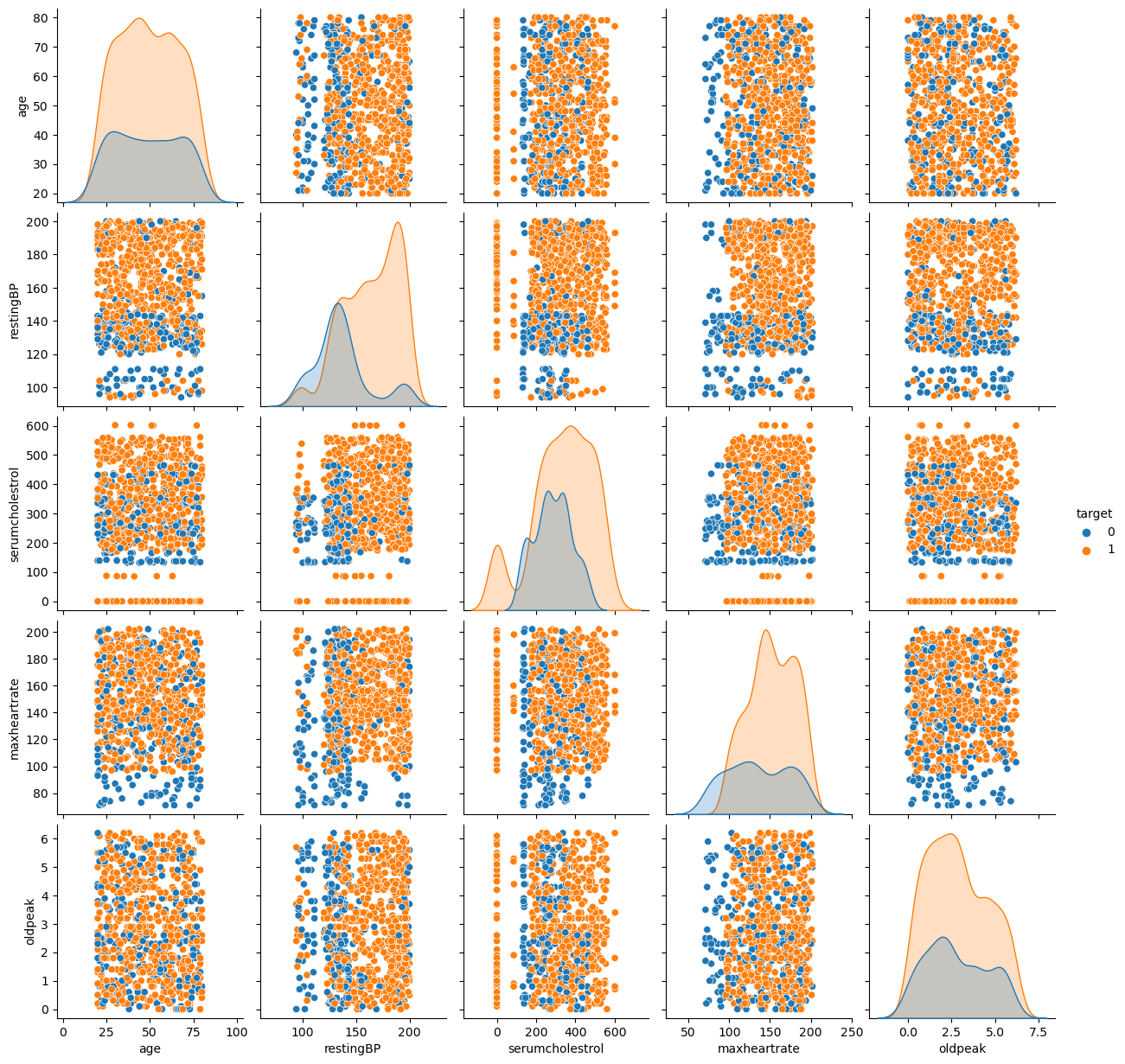
* + 1. **Bivariate Analysis:**

To find any correlations or linkages, pairwise relationships between the variables were investigated. Correlation matrices had to be created for this. According to the correlation matrix, the dataset's independent variables have correlation coefficients that are almost zero. One of the main presumptions of logistic regression analysis is met by this, which suggests a minimal chance of multicollinearity. Therefore, it seems that logistic regression is a good approach for this dataset.



* + 1. **Multivariate Analysis:**

The examination of the interplay among several variables revealed intricate patterns or trends. For this, we employed pair plotting.



**DATA ANALYSIS SOLUTION**

An 80:20 split of the data is made into training and testing sets. Having distinct datasets for training and assessing the models' performance is ensured by this phase. After the data was ready, we used three distinct algorithms—Random Forest, SVM, and Logistic Regression—to create classification models. We can determine which method is most suited for the task at hand by evaluating the performance of each model, which each has advantages and disadvantages.

**Random Forest**

The number of trees, the maximum depth of each tree, and the lowest number of samples needed to split a node using grid search with cross-validation were among the hyperparameters tested in our investigation. Superior prediction performance was demonstrated by the Random Forest model, which yielded a high overall accuracy of 97%.

**Support Vector Machine (SVM)**

Because SVM can use many kernel functions for decision-making, it is versatile and successful in high-dimensional spaces. Different hyperparameters, such as the regularization parameter C and the kernel selection (linear or radial basis function), were tested in our investigation. At 98% accuracy, the SVM model performed somewhat better than the Random Forest model.

**Logistic Regression**

We adjusted the regularization parameter C in our research to manage both underfitting and overfitting. The Logistic Regression model performed well, with an accuracy of 96%, which was marginally lower than the other two models.

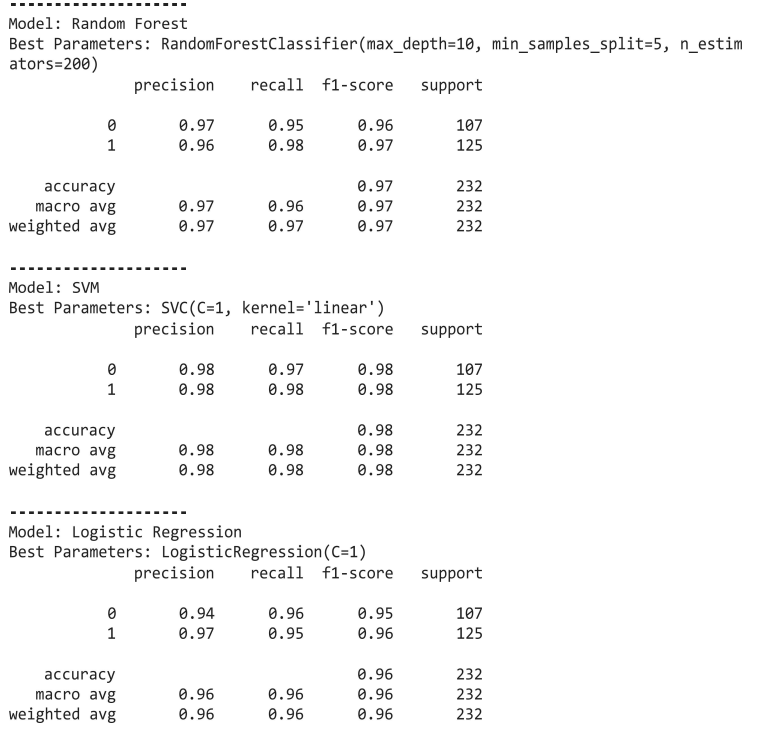
**Evaluation Metrics**

We utilized measures like precision, recall, F1-score, and total accuracy to assess how well the classification models performed. Recall is very important in our dataset since it highlights the true positive case detection, which is important in applications where the absence of positive cases can have dire repercussions. Recall offers a more focused knowledge of the model's performance in this area, which is particularly useful in medical environments like ours where the identification of individuals with heart disease is critical. It is a crucial indicator of the model's accuracy in identifying people who are at danger, reducing the possibility that important cases may go unnoticed.

**Insights and Comparison**

By examining the outcomes and observations from the classification report, we are able to make a number of inferences regarding the three models' respective performances. First of all, the three models' high accuracy shows how well they classified the target variable. While Logistic Regression fared somewhat worse but was still quite good, the SVM model showed the best accuracy, closely followed by the Random Forest model. Furthermore, looking at the models' recall, accuracy, and F1-score for both positive and negative classes shed light on how well the models can recognize examples of each class. In general, all three models demonstrated strong performance in categorizing the target variable; nevertheless, the optimal model selection is contingent upon the particular demands of the task, including interpretability, computing efficiency, and the significance of various assessment metrics.

* **Random Forest Model:**
  + Precision, recall, and F1-score for both classes (0 and 1) are quite high, indicating good performance.
  + The overall accuracy of the model is 0.97, which is very high.
  + The model performs well in classifying both positive and negative instances.
* **SVM Model:**
  + Precision, recall, and F1-score for both classes are also high, with an accuracy of 0.98.
  + The SVM model performs slightly better than the Random Forest model in terms of overall accuracy.
* **Logistic Regression Model:**
* Precision, recall, and F1-score are high for both classes, but slightly lower than the other two models.
* The overall accuracy of the Logistic Regression model is 0.96.



**CONCLUSION**

In summary, the classification study utilizing SVM, Random Forest, and Logistic Regression offered insightful information on the predictive modeling of the dataset. We were able to comprehend the capabilities of each algorithm by comparing their performance using evaluation criteria, since each algorithm showed both strengths and drawbacks. The SVM model demonstrated the highest accuracy, with Random Forest and Logistic Regression trailing closely after. Support Vector Machine (SVM) performed the best with recall of 98%. The properties of the dataset, the available processing power, and the needs of the task all play a role in selecting the optimal model. In summary, the present analysis underscores the significance of carefully assessing model performance and choosing the best classification method for the given task to arrive at well-informed conclusions.

**Advantages:**

* Model Performance: The Support Vector Machine (SVM) model had the greatest accuracy of 98% among the applied machine learning models, which showed good accuracy in predicting the existence of cardiovascular disease.
* Flexibility: The use of multiple algorithms allows for flexibility in model selection based on specific requirements and dataset characteristics.
* Interpretability: The coefficients from logistic regression are easy to understand, making it simpler to see how each factor affects the prediction.
* Scalability: Random Forest and SVM models can efficiently handle high-dimensional data and are scalable to bigger datasets.

**Limitations:**

* Imbalanced Dataset: The original dataset was imbalanced, with a significantly higher number of instances in one class compared to the other. This imbalance could potentially bias the model towards the majority class, affecting its performance.
* Computational Complexity: Random Forest, and SVM models, especially with grid search for hyperparameter tuning, can be computationally expensive, requiring more resources and time for training compared to Logistic Regression.
* Interpretability: While Logistic Regression provides easily interpretable coefficients, SVM and Random Forest models are less interpretable, making it challenging to understand the underlying decision-making process.
* Potential Overfitting: Without careful regularization and tuning, more complex models like Random Forest and SVM may be prone to overfitting, especially on smaller datasets.

**Operational Recommendations:**

* Address Imbalance: Further exploration of techniques to handle class imbalance, such as oversampling, undersampling, or using different evaluation metrics like F1-score, precision-recall curves, or ROC-AUC, is recommended to improve model performance.
* Optimize Computational Resources: Consideration of computational resources and time constraints should be made when selecting the final model. If computational resources are limited, Logistic Regression may be a more practical choice.
* Regularization: Implement regularization techniques such as L1 or L2 regularization for SVM and Random Forest models to mitigate overfitting and improve generalization performance.
* Model Interpretation: While SVM and Random Forest models provide high predictive performance, efforts should be made to interpret their decisions through techniques like feature importance analysis, partial dependence plots, or SHAP values to enhance model transparency and trust.
* Continuous Monitoring: Regular monitoring and validation of the deployed model's performance on new data are essential to ensure its continued effectiveness and relevance in real-world scenarios.

**Appendix**

