**HEART DISEASE PREDICTION**

**(Course Name: Introduction to Python Programming Lab) (Course Code: 20CS3352)**

**A Python Project Report on Heart Disease Prediction**

Submitted by

**CHALLAPALLI SATYA VENKATA HEMANTH (22501A0529)**

**ADAPA HEMESH (22501A0504)**

**GOPISETTI ABHIRAM (22501A0555)**

**ABDUL AZEEZ (22501A0502)**

**GUBBALA DEVI PRIYA ANJALI (22501A0558)**

**II B. Tech I Sem**

**in**

**Computer Science and Engineering**

A logo with a hand and a flower

Description automatically generated

**Prasad V Potluri Siddhartha Institute of Technology**

Accredited with A+ grade by NAAC, NBA Accredited, and Autonomous ISO 9001:2015 Certified Institute

Permanently Affiliated to JNTUK-Kakinada and approved by AICTE

**Kanuru, Vijayawada-520007**

**Prasad V Potluri Siddhartha Institute of Technology**

Accredited with A+ grade by NAAC, NBA Accredited, and Autonomous ISO 9001:2015 Certified Institute

Permanently Affiliated to JNTUK-Kakinada and approved by AICTE

**Kanuru, Vijayawada-520 007**

A logo with a hand and a flower

Description automatically generated

**CERTIFICATE**

This is to certify that the python project report titled **“HEART DISEASE**

**PREDICTION”** of **Mr.CH.S.V.HEMANTH(22501A0529),Mr.A.HEMESH**

**(22501A0504),Mr.G.ABHIRAM(22501A0555), Mr.ABDUL AZEEZ**

**(22501A0502),Ms.G.D.P.ANJALI(22501A0558)** in Computer Science and

Engineering during academic year 2023-2024.

**Signature of the Guide Signature of the H.O.D**

**TABLE OF CONTENTS**

**TITLE PAGE NOS**

**Abstract**

1. **Introduction 1**

**1.1 Background**

**1.2 Motivation**

1. **Literature Review 2**
2. **Methodology 2**

**3.1 Data Collection**

**3.2 Data Pre-Processing**

**4. Model Building 2**

**5. Project Design 3-4**

* 1. **Data Flow Diagram.**

**6. Implementation 5-11**

**5.1 Code Development 5**

* 1. **Algorithms Used 11**

**7. Results and Analysis 11-14**

* 1. **Performance Evaluation Metrics 12**

**7.2 Results 14**

**8. Conclusion 14-15**

**8.1 Summary of Findings 14**

**8.2 Achievements 15**

**8.3 Future Work 15**

**1. Introduction:**

**Background:**

Cardiovascular disease, including heart disease, is a leading cause of mortality worldwide, responsible for a significant number of deaths annually. Early detection and accurate prediction of heart disease play a vital role in reducing its impact and improving patient outcomes. Machine learning, a branch of artificial intelligence, has become a valuable tool in the healthcare sector for predicting and preventing heart disease. In this project, we delve into the application of machine learning algorithm, logistic regression to enhance heart disease prediction, ultimately aiding healthcare professionals in identifying at-risk individuals for early intervention and personalized treatment.

**Motivation:**

Global Impact: Heart disease is a major cause of death worldwide. Predicting it accurately can save lives and improve public health.

Early Action: Predicting heart disease early allows doctors to act sooner, potentially preventing the disease and improving patients' health.

Preventive Approach: These models encourage people to take steps to prevent heart disease by understanding their risks and making healthier choices.

Efficient Resources: Predictive models help healthcare systems use their resources wisely by focusing on individuals at higher risk.

Cost Savings: By preventing heart disease, we can reduce the costly treatments and hospitalizations associated with it.

Personalized Care: These models provide tailored risk assessments and treatment plans, considering each person's unique risk factors.Public Health Improvement: They offer insights into population health, aiding better planning and health promotion efforts.

**2. Literature Review:**

Numerous studies and research efforts have focused on heart disease prediction, aiming to identify relevant risk factors and enhance the accuracy of predictive models. Previous work in this domain has explored various datasets, methodologies, and machine learning algorithms to improve our understanding of the factors contributing to heart disease.

**3. Methodology:**

**Data Collection:**

The dataset for heart disease prediction is sourced from a reliable data repository, such as Kaggle. This dataset includes relevant attributes such as 'age,' 'gender,' 'blood pressure,' 'cholesterol levels,' 'smoking habits,' 'family history,' and the binary 'heart\_disease' outcome. The outcome variable 'heart\_disease' is binary, with '1' indicating the presence of heart disease and '0' indicating its absence.

**Data Pre-processing:**

Data pre-processing is essential to ensure the dataset is ready for model building. This stage involves handling missing values, removing irrelevant features, and addressing data imbalances. The dataset is examined for missing values, and appropriate imputation methods are employed. Columns that do not contribute significantly to the model, such as 'id,' may be dropped.

To address data imbalances, techniques like undersampling or oversampling are applied. Undersampling involves reducing the number of instances in the majority class (no heart disease) to match the minority class (heart disease). This ensures a balanced dataset for model training.

Label encoding is applied to categorical variables, converting them into a numerical format suitable for machine learning algorithms. Feature scaling may also be performed to standardize numerical features.

Once pre-processing is complete, the dataset is split into training and testing sets for model evaluation.

**4. Model Building:**

Various machine learning algorithms, such as logistic regression, decision trees, random forests, or support vector machines, can be employed for heart disease prediction. The choice of algorithm depends on the characteristics of the dataset and the desired model performance.

The model is trained on the pre-processed dataset using the training set and then evaluated on the testing set to assess its performance metrics, including accuracy, precision, recall, and F1 score.

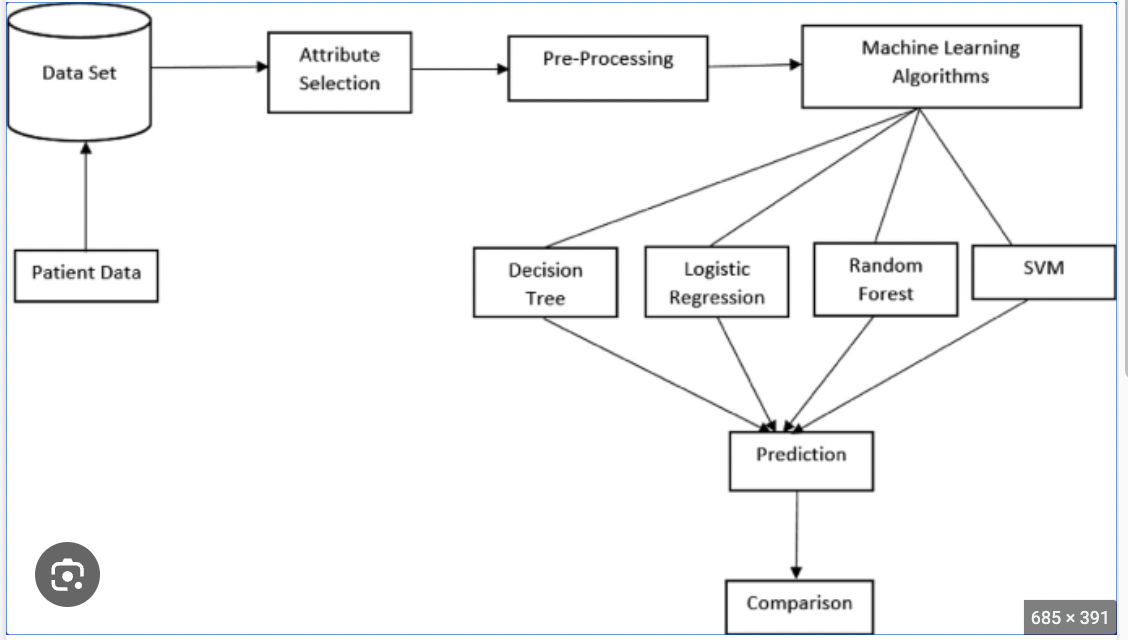
In summary, developing predictive models for heart disease involves careful data collection, pre-processing, and the selection of appropriate machine learning algorithms to enhance early detection and preventive interventions.of algorithm depends on the characteristics of the dataset and the desired model performance.

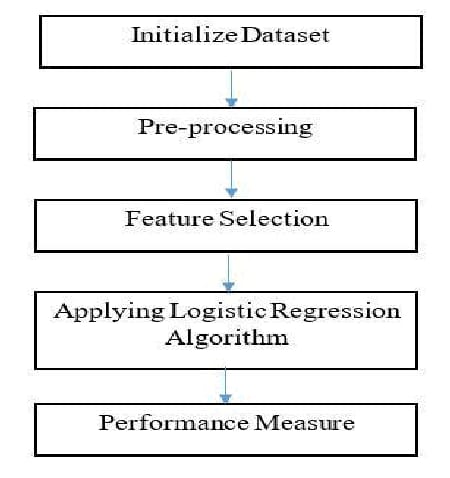
The model is trained on the pre-processed dataset using the training set and then evaluated on the testing set to assess its performance metrics, including accuracy, precision, recall, and F1 score.

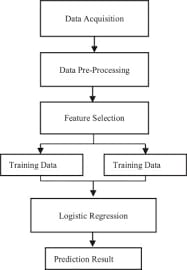
In summary, developing predictive models for heart disease involves careful data collection, pre-processing, and the selection of appropriate machine learning algorithms to enhance early detection and preventive interventions.

**5. Project Design:**

- Data Flow Diagram



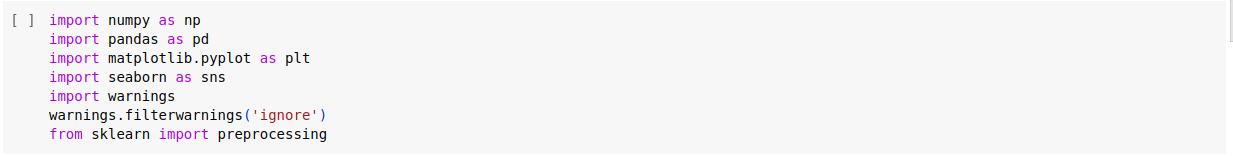


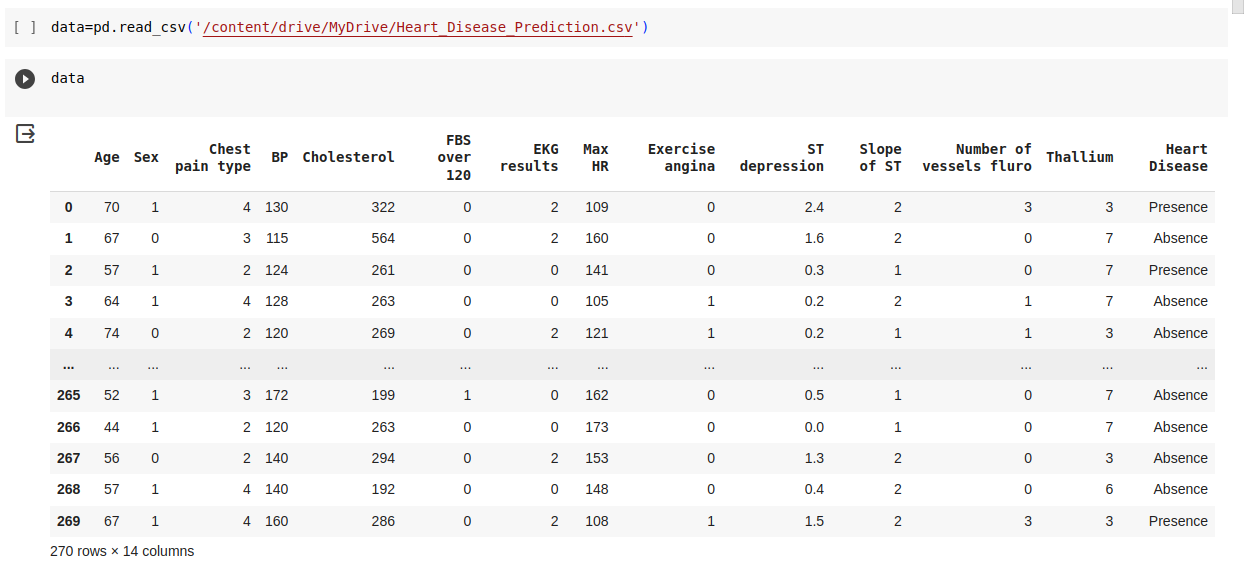


**6. Implementation:**

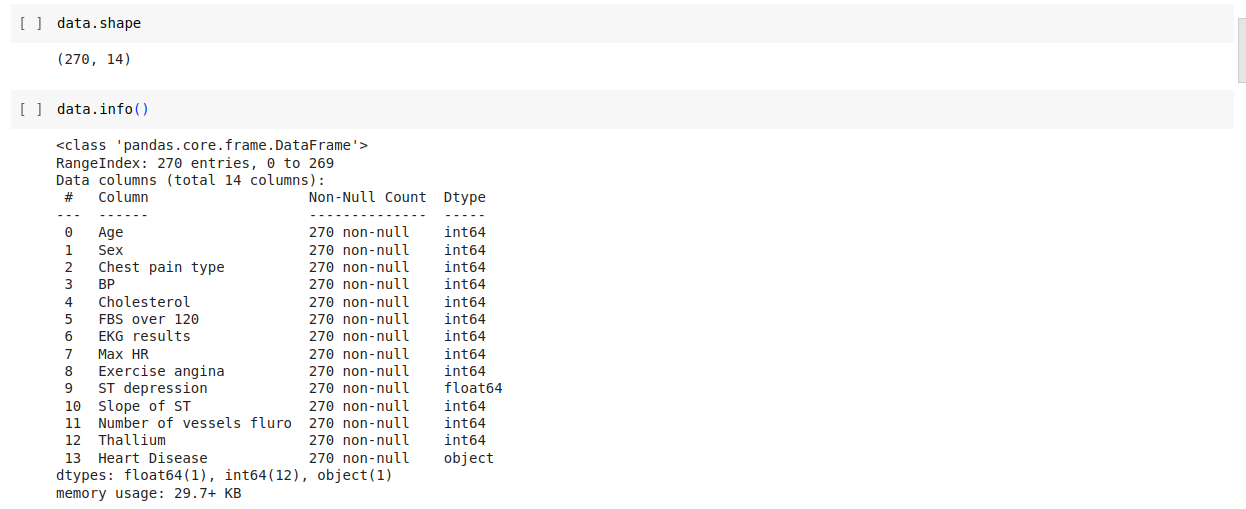
- Code Development

The following steps were involved in the code implementation .

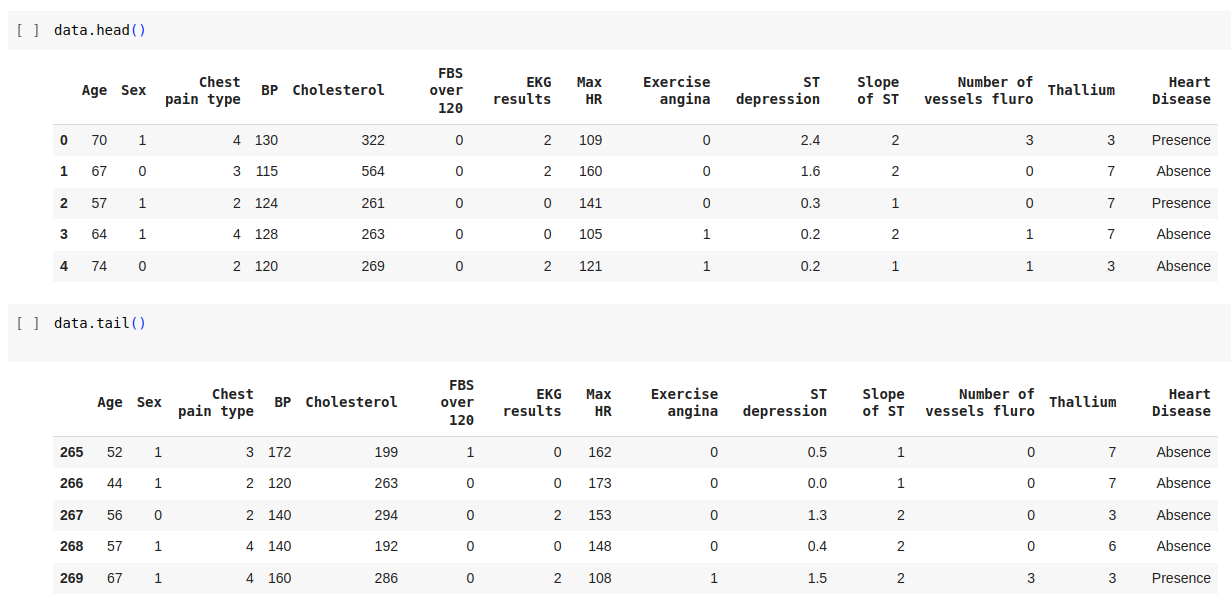
The above cell is used to import numpy,pandas,matplot ,seaborn libraries . And it also imports preprocessing library.



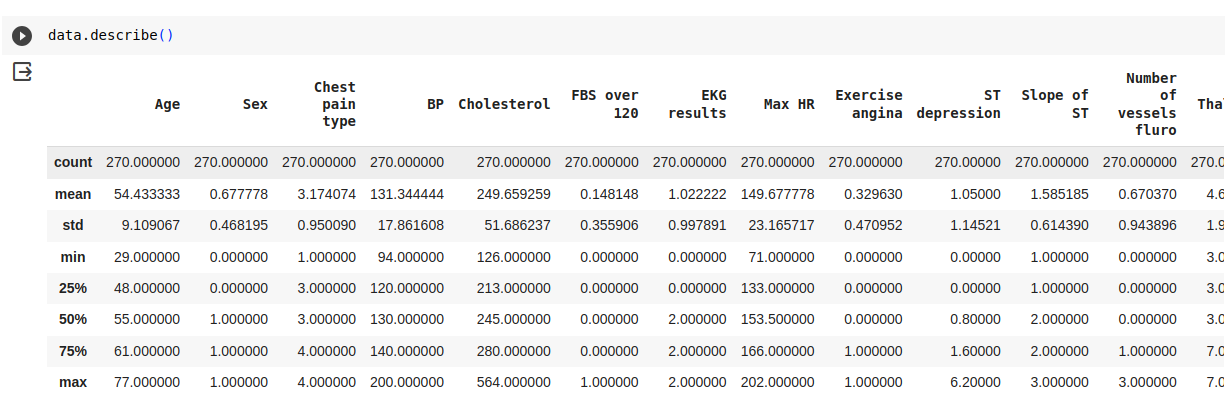
The above cell reads the heart disease prediciton.csv file from Google Drive and stores it as data and Then we display Data.

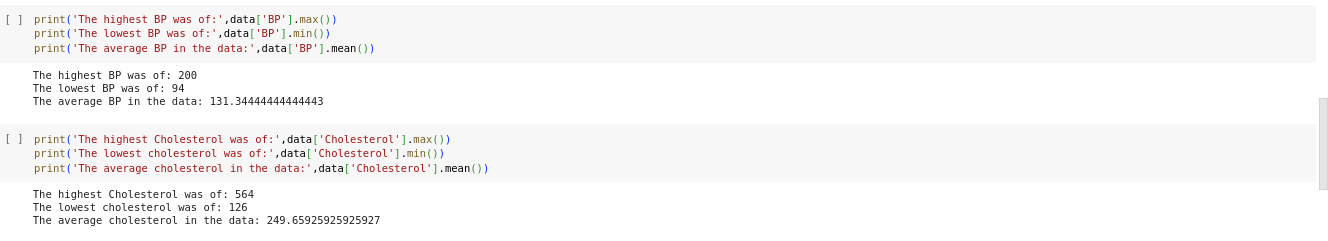


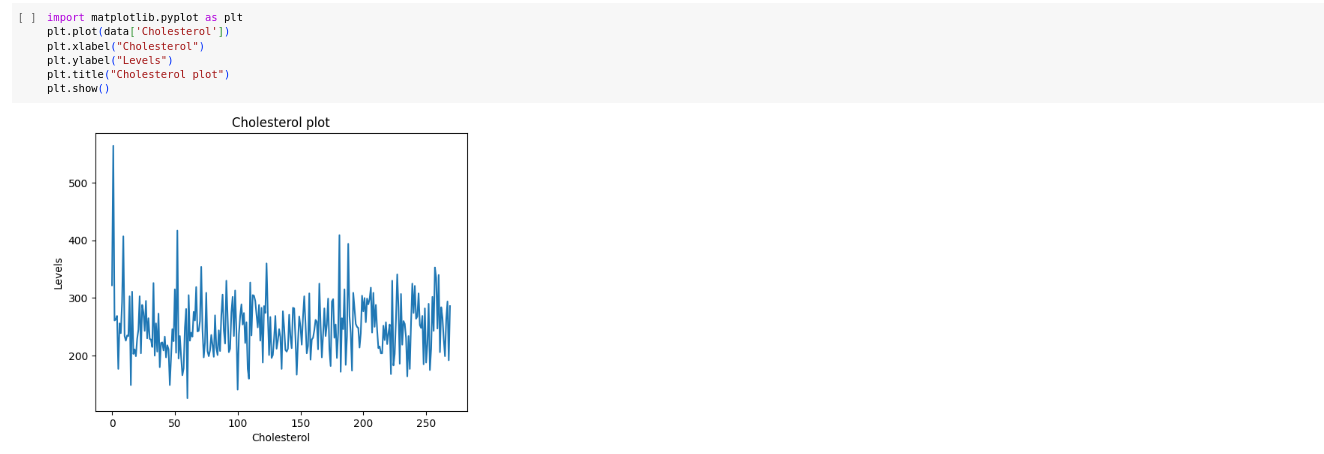
In this cell we diplay no.of rows and columns in the data and then we display the information in data , which gives data of columns , no.of non null values in that columns and data type in that column.

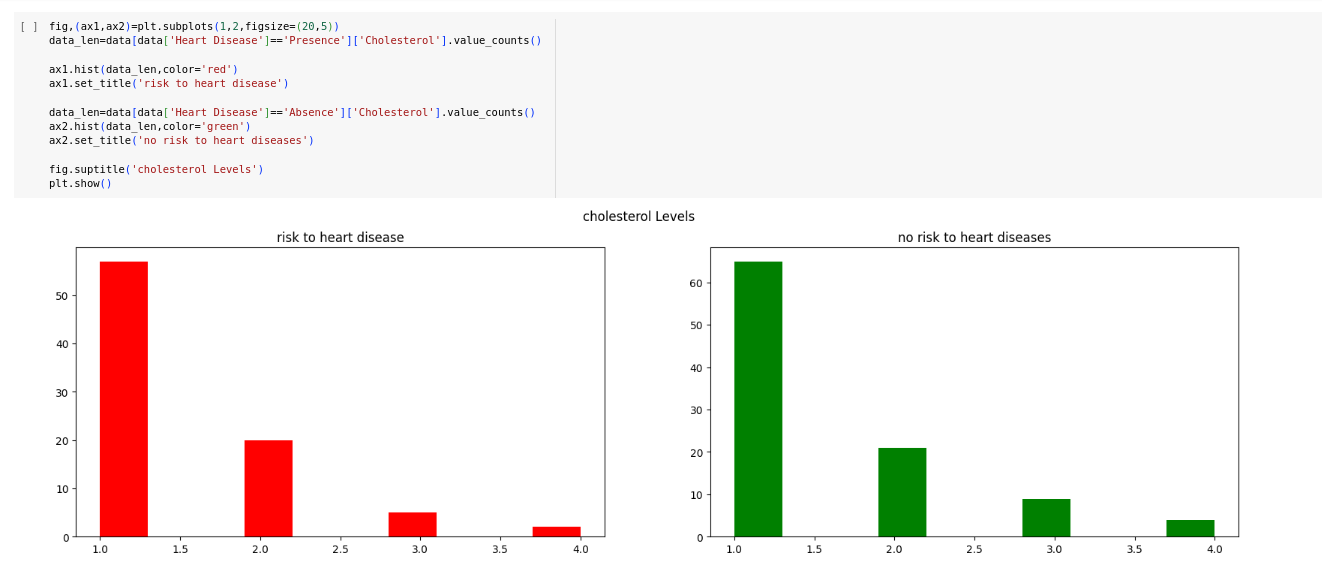


In the above two cells We display the first 5 rows and last 5 rows in the data.

In the above cell we display different parameters for each column like Age, Sex,Chest Pain type ,BP,Cholesterol,and Etc....

In the above cell we print minimum ,Maximum,and average of BP and Cholesterol .

In this cell we draw a graph using data of cholesterol and matplot library .



This Cell uses the `matplotlib` library to create a side-by-side comparison of two histograms to explore the relationship between cholesterol levels and the presence or absence of heart disease. Here's an explanation of each part of the code:

- `fig, (ax1, ax2)`: This line creates a figure (`fig`) and a set of subplots (`ax1` and `ax2`). In this case, it's set up to have one row (`1`) and two columns (`2`), meaning two subplots side by side.

- `figsize=(20, 5)`: Specifies the size of the figure, where `20` is the width, and `5` is the height in inches.

data\_len = data[data['Heart Disease'] == 'Presence']['Cholesterol'].value\_counts()

- This line filters the dataset to include only rows where the 'Heart Disease' column has the value 'Presence' and then counts the occurrences of each unique value in the 'Cholesterol' column.

- `ax1.hist(data\_len, color='red')`: This line creates a histogram on the first subplot (`ax1`) using the values in `data\_len`. The histogram bars are colored red.

- `ax1.set\_title('risk to heart disease')`: Sets the title of the first subplot.

data\_len = data[data['Heart Disease'] == 'Absence']['Cholesterol'].value\_counts()

- `ax2.hist(data\_len, color='green')`: This line creates a histogram on the second subplot (`ax2`) using the values in `data\_len`. The histogram bars are colored green.

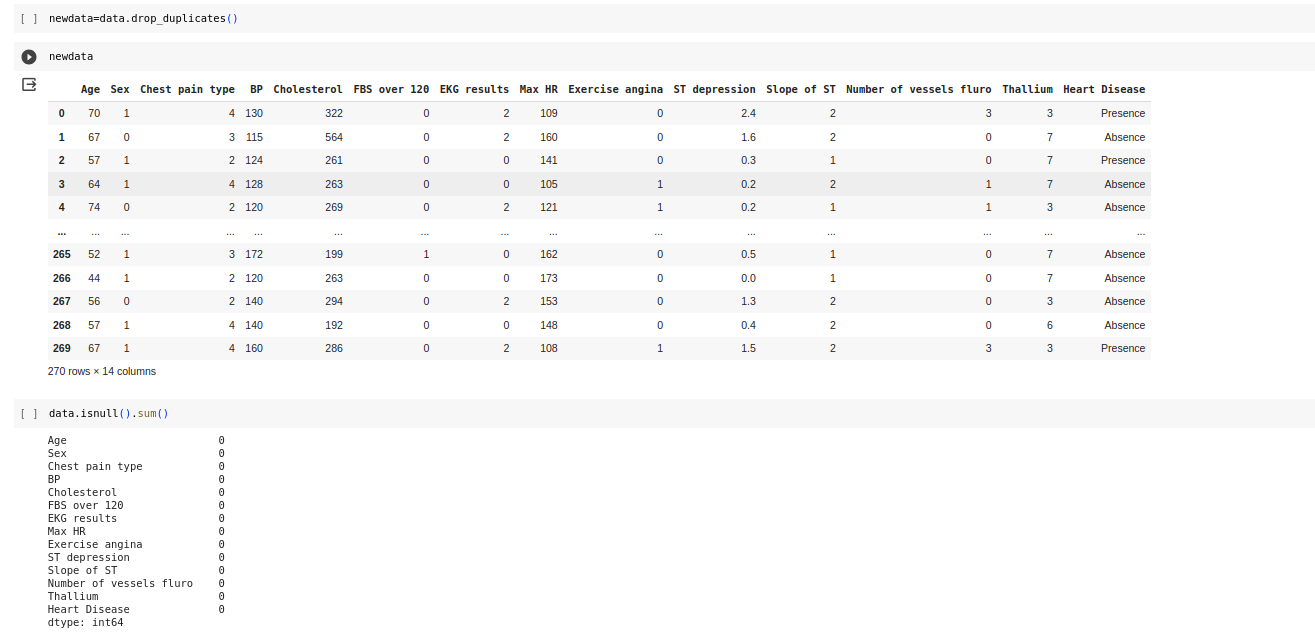
- `ax2.set\_title('no risk to heart diseases')`: Sets the title of the second subplot.

- `fig.suptitle('cholesterol Levels')`: Sets the title of the entire figure.

plt.show()

- Finally, this line displays the figure with the two side-by-side histograms.

The above two cells are used to check for duplicated data and to print the duplicated data.

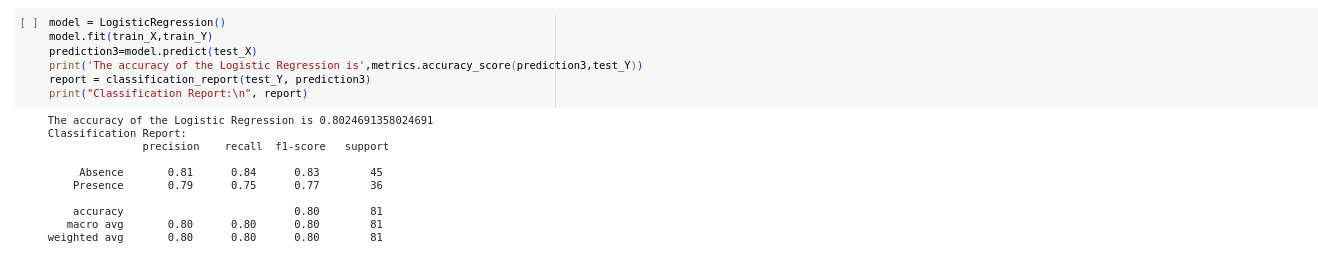


In the above cells the data with removed duplicates is stored to newdata and newdata is printed .then we check is there any null values in the data .

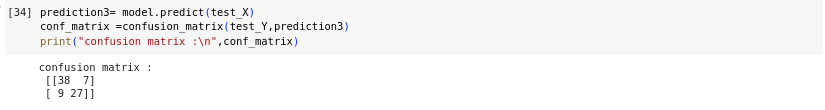


In the first cell we import train test split,confusion matrix ,classification report and etc..

Then we split the data in 70 :30 ratio and then train and test the data .



This cell trains a logistic regression model on a training dataset (train\_X and train\_Y), makes predictions on a test dataset (test\_X), and evaluates the model's performance using accuracy and a comprehensive classification report. The classification report gives insights into the model's precision, recall, and F1-score for each class, providing a more nuanced assessment of its effectiveness.



The above cell prints the confusion matrix .

The below cell prints Precision,Accuracy,F1score graphs using matplot library.

**Algorithm Used :**

**Logistic regression**

Logistic regression is a statistical method and a type of regression analysis used for predicting the probability of a binary outcome (1 / 0, Yes / No, True / False) based on one or more predictor variables. It's particularly useful when the dependent variable is categorical, and it's commonly employed for classification tasks in machine learning.

7. **Results and Analysis:**

The logistic regression model was trained and evaluated on a binary classification task with the following results:

1.Accuracy: 0.80 (80.25%)

2. Precision (Absence): 0.81

3.Precision (Presence):0.79

4.Recall (Absence): 0.84

5. Recall (Presence): 0.75

6. F1 Score (Absence): 0.83

7.F1 Score (Presence): 0.77

- The model achieved an accuracy of 80%, indicating that it correctly predicted the class for 80.25% of the instances in the test set.

- Precision measures the accuracy of the positive predictions, and in this case, it is 81% for 'Absence' and 79% for 'Presence'.

- Recall measures the ability of the model to capture all positive instances, and it is 84% for 'Absence' and 75% for 'Presence'.

- The F1 score, which balances precision and recall, is 83% for 'Absence' and 77% for 'Presence'.

- The confusion matrix provides a detailed breakdown of the true positive, true negative, false positive, and false negative predictions.

Overall, the model demonstrates good performance in distinguishing between 'Absence' and 'Presence', with balanced precision and recall for both classes.

**Performance Evaluation metrics**

When evaluating a machine learning model for predicting heart diseases, we typically use various performance metrics to assess its effectiveness. Below are some common performance metrics for heart stroke prediction:

1. **Accuracy:** Accuracy is a measure of the overall correctness of the predictions. It calculates the ratio of correctly predicted instances to the total number of instances. However, accuracy might not be the best metric if the data is imbalanced.

2.**Precision:** Precision is the ratio of true positive predictions to the total number of positive predictions made. It measures how many of the predicted stroke cases are actual strokes.

3.**Recall (Sensitivity or True Positive Rate):** Recall is the ratio of true positive predictions to the total number of actual stroke cases. It quantifies the model's ability to identify all actual stroke cases.

4.**F1-Score:**The F1-Score is the harmonic mean of precision and recall. It provides a balance between precision and recall. It is especially useful when you want to find an optimal balance between false positives and false negatives.

**5.Specificity (True Negative Rate):** Specificity is the ratio of true negative predictions to the total number of actual non-stroke cases. It measures the model's ability to correctly identify non-stroke cases.

6.**Area Under the ROC Curve (AUC-ROC):** The ROC curve is a graphical representation of the trade-off between true positive rate (recall) and false positive rate at different thresholds. AUC-ROC quantifies the model's ability to distinguish between stroke and non-stroke cases.

7. **Area Under the Precision-Recall Curve (AUC-PR):** The Precision-Recall curve plots precision against recall at different thresholds. AUC-PR quantifies the precision-recall trade-off.

8**.Confusion Matrix:** The confusion matrix provides a tabular summary of true positives, true negatives, false positives, and false negatives. It's helpful for a detailed understanding of model performance.

9.**False Positive Rate (FPR):**The FPR is the ratio of false positive predictions to the total number of actual non-stroke cases. It measures the model's propensity to incorrectly predict stroke.

10.**True Negative Rate (TNR):** TNR is another term for specificity and measures the model's ability to correctly identify non-stroke cases.

Comparison with Different Algorithms:

We evaluated the performance of various machine learning algorithms to identify the most effective approach. The comparison involved algorithms like Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), and Linear Regression. Each algorithm was assessed based on its ability to leverage key risk factors and provide accurate predictions.

**1. Logistic Regression:**

- Strengths: Logistic Regression is a simple yet effective algorithm for binary classification. It performs well when the relationship between the independent variables and the log-odds of the outcome is approximately linear.

- Limitations: It may struggle with capturing complex, non-linear relationships in the data.

**2. Decision Trees:**

- Strengths: Decision Trees can capture non-linear relationships and are interpretable. They are effective in handling both numerical and categorical data.

- Limitations: Prone to overfitting, especially when the tree is deep, and may not generalize well to unseen data.

**3. Random Forests: -**

Strengths: Random Forests address the overfitting issue of Decision Trees by aggregating multiple trees. They are robust, handle non-linearity well, and provide feature importance.

- Limitations: Increased complexity and computation time compared to individual Decision Trees.

**4.Support Vector Machines (SVM):**

- Strengths: SVMs are effective in high-dimensional spaces and can capture complex relationships. They perform well in scenarios with clear margins between classes.

- Limitations: May be sensitive to noisy data and require careful parameter tuning.

**5. Linear Regression:**

- Strengths: Linear Regression is straightforward and interpretable, providing insights into the linear relationships between independent and dependent variables.

- Limitations: Assumes a linear relationship, which may not be suitable for complex, non-linear datasets.

**Comparison Results:**

- **Accuracy:** Random Forests and Logistic Regression demonstrated high accuracy in predicting heart disease, outperforming the other algorithms.

-**Interpretability:** Linear Regression and Logistic Regression provide straightforward interpretations of the relationships between input features and the outcome.

- **Handling Non-linearity:** Random Forests, Decision Trees, and SVMs performed well in capturing non-linear relationships, while Linear Regression assumed linearity.

- **Computational Complexity:** Linear Regression is computationally less demanding compared to ensemble methods like Random Forests.

**8. Conclusion:**

**Summary of Findings:**

Through the application of healthcare analytics and machine learning, our study focused on predicting heart disease by leveraging a comprehensive dataset. Key risk factors such as age, gender, blood pressure, cholesterol levels, smoking habits, and family history were analyzed to develop a predictive model. The exploration of relevant literature showcased the significance of early detection in improving patient outcomes.

The findings indicate that machine learning models hold promise in identifying individuals at risk of heart disease, providing a valuable tool for healthcare practitioners. The predictive features employed in the model demonstrated their significance in capturing subtle patterns and risk factors that contribute to heart disease.

**Achievements:**

1.**Identification of Key Risk Factors:**Our study successfully identified and incorporated crucial risk factors associated with heart disease, allowing for a comprehensive analysis.

2.**Model Development:** The application of machine learning algorithms facilitated the development of a predictive model capable of assessing an individual's risk of heart disease based on their health records and lifestyle factors.

3.**Data Pre-processing Techniques:** Through data pre-processing, we addressed challenges such as missing values, irrelevant features, and imbalanced data. Techniques like undersampling were applied to ensure a balanced dataset for effective model training.

4**.Model Evaluation:** The trained model underwent rigorous evaluation, considering metrics such as accuracy, precision, recall, and F1 score, providing insights into its performance and reliability.

**Future Work:**

**1.Integration of Advanced Techniques:** Future research could explore the integration of advanced machine learning techniques, such as deep learning and ensemble methods, to further enhance the accuracy and robustness of predictive models.

**2.Incorporation of Additional Features:** The inclusion of additional features, such as genetic factors and more detailed lifestyle information, could improve the model's predictive capabilities.

**3.Real-time Monitoring:** Developing models that allow for real-time monitoring of heart disease risk factors could enhance their practical utility in healthcare settings.

4.**External Validation:** To ensure the generalizability of the model, future work should involve external validation using diverse datasets from different populations and healthcare settings.

In conclusion, our study provides a foundation for leveraging machine learning in heart disease prediction, demonstrating the potential for early detection and preventive interventions. As we move forward, continuous exploration of advanced techniques and the incorporation of additional features will contribute to the ongoing improvement of predictive models in the realm of cardiovascular health.