Cardiovascular Disease Risk Prediction

INTRODUCTION:

In recent years, the rapid growth of healthcare data and advancements in machine learning have created new opportunities for improving disease prediction and preventive care. Machine learning models can process large volumes of medical data and uncover hidden patterns that may not be easily identifiable through traditional analysis.

By applying these techniques to cardiovascular health data, this project aims to enhance risk assessment accuracy, reduce diagnostic delays, and provide a scalable solution that can assist medical professionals in monitoring patient health more effectively. Such predictive systems play a crucial role in promoting early intervention and improving long-term patient outcomes.

The Cardiovascular Disease Risk Prediction Dataset on Kaggle is a structured health dataset designed for building and evaluating models that estimate the likelihood of cardiovascular disease in individuals. It contains **comprehensive synthetic healthcare data** with multiple predictors related to personal, biological, and lifestyle risk factors for heart-related illness.

According to the dataset page description, it comprises **19 variables**, including both numerical and categorical features, which capture key health indicators and demographic information relevant for assessing cardiovascular risk in patients.

Each row in the dataset represents an individual, and the columns correspond to different health attributes such as age, blood pressure, cholesterol levels, body measurements, and lifestyle factors (e.g., smoking, physical activity). Together these features serve as input variables in predictive models that aim to **classify whether a person is at risk of developing a cardiovascular condition**—typically framed as a supervised binary classification problem.

This dataset is widely used in data science and machine learning projects for **exploratory data analysis (EDA)**, feature engineering, and model building, helping researchers and students explore how various factors influence cardiovascular disease onset. It provides a realistic benchmark for comparing the performance of different machine learning algorithms in predicting cardiovascular risk, making it suitable for both academic and practical applications in healthcare analytics.

AIM OF THE PROJECT:

The aim of this project is to **develop a machine learning–based model to predict the risk of cardiovascular disease** in individuals using clinical, demographic, and lifestyle-related data. By analyzing key health factors such as age, blood pressure, cholesterol levels, body measurements, and habits like smoking and physical activity, the project seeks to identify patterns that contribute to cardiovascular disease.

The objective is to **build an accurate and reliable predictive system** that can classify individuals as being at risk or not at risk of cardiovascular disease. This helps in **early detection and prevention**, enabling timely medical intervention and improved healthcare decision-making. Additionally, the project evaluates the performance of different machine learning algorithms to determine the most effective model for cardiovascular risk prediction.

PROBLEM STATEMENT:

Cardiovascular diseases pose a serious global health challenge due to their high prevalence and mortality rate. Many risk factors such as age, blood pressure, cholesterol levels, obesity, smoking habits, and physical inactivity contribute to the development of heart disease, but identifying high-risk individuals at an early stage remains difficult. Conventional diagnostic approaches often require extensive medical tests and expert interpretation, which may delay timely detection.

To address this issue, there is a need for an **efficient predictive system** that can analyze patient health data and accurately assess cardiovascular disease risk. This project focuses on using machine learning techniques to process clinical and lifestyle data to predict the likelihood of cardiovascular disease, enabling early intervention, better preventive care, and improved decision support for healthcare professionals.

Project Workflow – Key Points

- Collected cardiovascular disease dataset from Kaggle
- Loaded the dataset using Python libraries
- Performed data inspection to understand structure and features
- Cleaned the data by checking missing values and duplicates
- Encoded categorical variables into numerical format
- Scaled numerical features for uniformity
- Conducted Exploratory Data Analysis (EDA) using statistical methods and visualizations
- Identified important features influencing cardiovascular disease risk
- Split the dataset into training and testing sets
- Implemented supervised machine learning model
- Trained the model using training data
- Evaluated model performance using accuracy and other metrics
- Analyzed results and drew conclusions

DATA UNDERSTANDING:

To gain deeper insights into the dataset, various inspection methods were use :

1. Dataset information was checked to identify numerical and categorical attributes.
2. The number of unique values in each column was analyzed to understand feature variability.
3. Unique values of categorical attributes such as **Gender** were examined to understand encoding.
4. Missing values were checked across all columns, and it was observed that the dataset contained no significant missing or duplicate values.

DATA CLEANING:

After importing all libraries and Data cleaning is an important step to ensure the accuracy and reliability of the dataset before applying machine learning algorithms. In this project, data cleaning was performed to remove inconsistencies and prepare the dataset for further analysis.

First, the dataset structure was examined to understand the number of rows, columns, and data types. This helped in identifying numerical and categorical attributes present in the dataset.

Next, the dataset was checked for missing values across all columns. Missing values can negatively impact model performance. The results showed that there were no significant missing values in the dataset, ensuring data completeness.

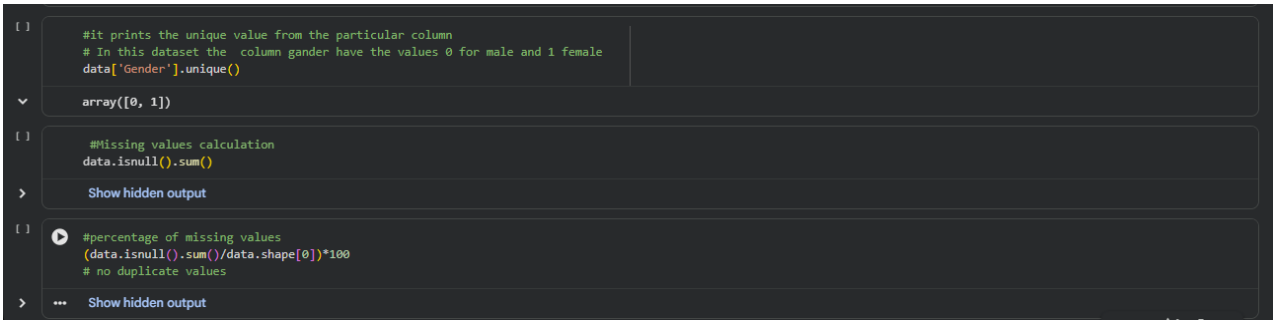


fig: - Finding Missing values and Unique values

The dataset was then checked for duplicate records. Duplicate entries can cause biased learning by repeating the same information. No duplicate records were found, confirming that each record represents a unique individual.

In this dataset, The Values were stored as numeric already and we can easily implement to the Machine Learning Models. We no need to encode the values.

Finally, numerical features were reviewed to ensure consistency and correctness. After completing these steps, the dataset was clean, structured, and ready for further processing and model implementation.

DATA PREPROCESSING:

Data preprocessing involved separating features and target variables, encoding categorical data, splitting the dataset into training and testing sets, and scaling numerical features to prepare the data for machine learning models.

1. Checking Unique Values in Gender Column

The first code cell checks the unique values present in the **Gender** column. This step is performed to understand how gender is encoded in the dataset. The output shows that gender is already

represented numerically, where **0 indicates male** and **1 indicates female**. This verification confirms that no additional encoding is required for this categorical variable.

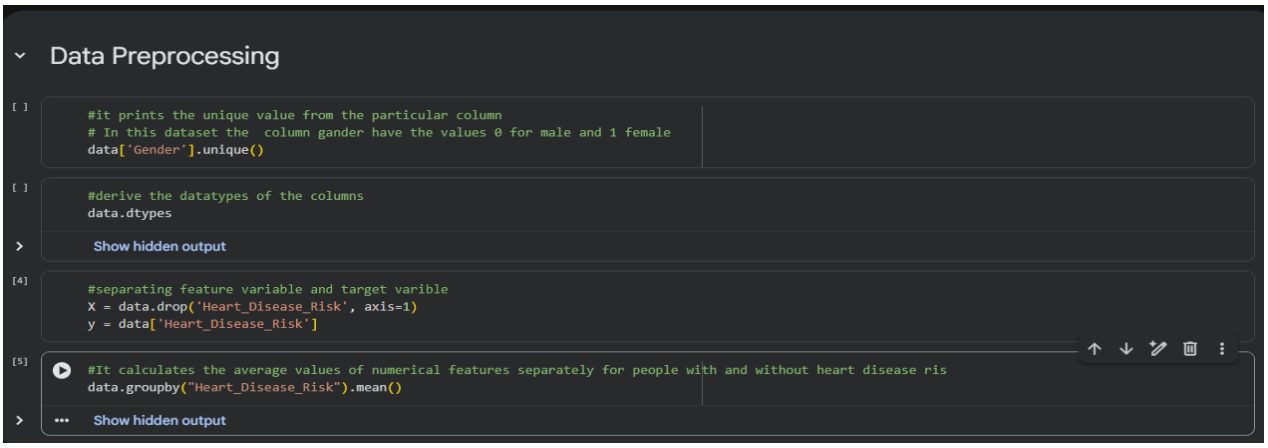


fig: - Data Preprocessing

2. Identifying Data Types of Columns

The next step displays the data types of all columns in the dataset. This helps distinguish between numerical and categorical features and ensures that all variables are in the correct format for machine learning algorithms. Verifying data types is an important preprocessing step to avoid errors during model training.

3. Separating Features and Target Variable

In this step, the dataset is divided into **independent variables (X)** and the **dependent variable (y)**.X contains all predictor features except the target column, y contains the target variable **Heart_Disease_Risk**. This separation is essential for supervised machine learning, as it allows the model to learn from input features and predict the target outcome.

4. Grouping Data by Target Variable

The final step groups the dataset based on the **Heart_Disease_Risk** column and calculates the mean of numerical features for each group. This helps in comparing average health indicators between individuals **with heart disease risk** and **without heart disease risk**. Although this step is part of exploratory analysis, it also supports preprocessing by providing insights into feature importance.

STATISTICAL IMPLEMENTATION:

In this project, statistical methods were implemented to understand the distribution of data and analyze differences between individuals with and without cardiovascular disease risk.First, descriptive statistical measures were computed using built-in functions. These statistics include mean, minimum, maximum, and standard deviation for numerical features such as age, blood pressure, cholesterol level, and other health indicators. This step helped in understanding the overall distribution and variability of the dataset.

Next, group-wise statistical analysis was performed by grouping the dataset based on the target variable **Heart_Disease_Risk**. Using the `group by ()` operation along with the `mean ()` function, the average values of numerical features were calculated separately for individuals with heart disease risk and those without risk. This comparison helped identify how health indicators differ between two groups.

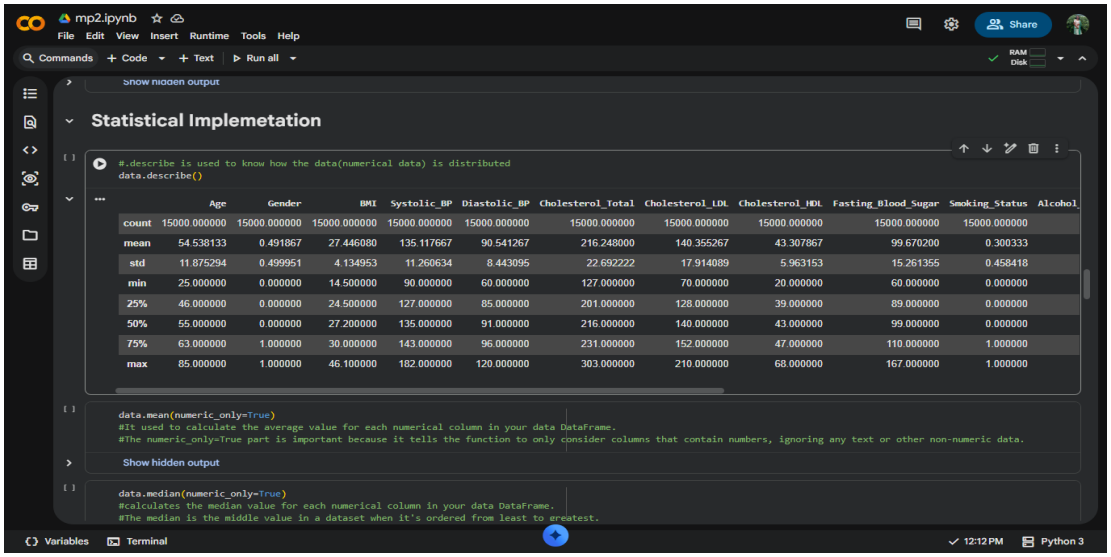


fig: - Statistical implementation

These statistical observations provided important insights into the dataset and supported further exploratory data analysis and model development. Overall, statistical implementation helped summarize the dataset quantitatively, identify key patterns, and validate assumptions before applying machine learning models.

EXPLORATORY DATA ANALYSIS:

Exploratory Data Analysis (EDA) was performed to visually and statistically understand patterns, trends, and relationships within the cardiovascular disease dataset. EDA helps in identifying important features and understanding how they influence the target variable before model training

1.Univariate Analysis:

Univariate analysis was carried out to study the distribution of individual variables. Histograms were used to visualize numerical features such as age, cholesterol level, and blood pressure. This analysis helped in understanding data distribution, range of values, and the presence of skewness or extreme values.

```
#histogram
data.hist(figsize=(10,10))
```

Histograms were plotted for all numerical features to analyze their data distribution. This visualization helps in identifying the range, frequency, skewness, and presence of outliers in each feature, providing a better understanding of the dataset before model training.

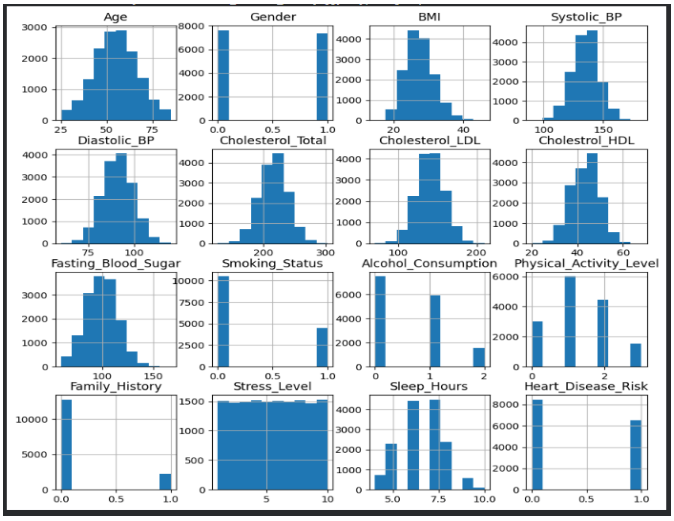
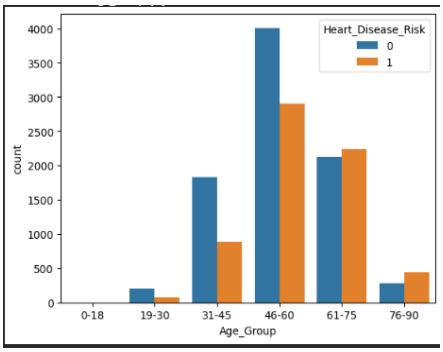
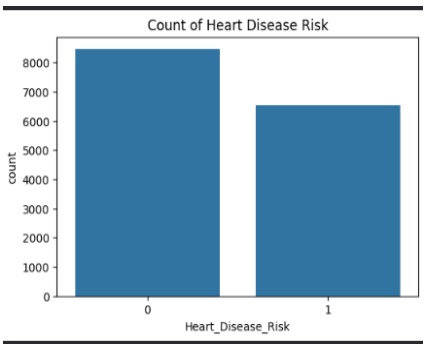


fig: - Histogram

2.Bivariate Analysis:

Bivariate analysis was performed to examine the relationship between input features and the target variable **Heart_Disease_Risk**. Count plots were used to compare the distribution of cardiovascular risk across categorical variables such as gender. This helped in identifying differences between risk and non-risk groups. This is a count plot (implemented as a bar plot) showing counts of categories.





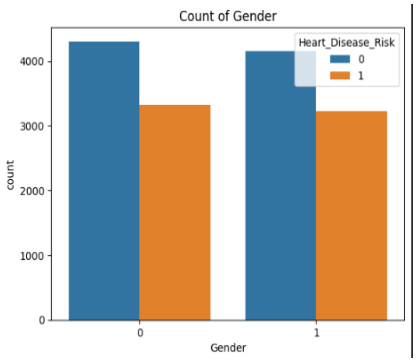


fig: - Bivariate Analysis

3.Multivariate Analysis:

To analyze relationships between multiple numerical features, a correlation matrix and heatmap were created. This helped in identifying strongly correlated variables and understanding how different health indicators are related to each other and to cardiovascular disease risk.

Heatmap:

A heatmap is used to visualize the correlation between features, with colors representing the strength of relationships.

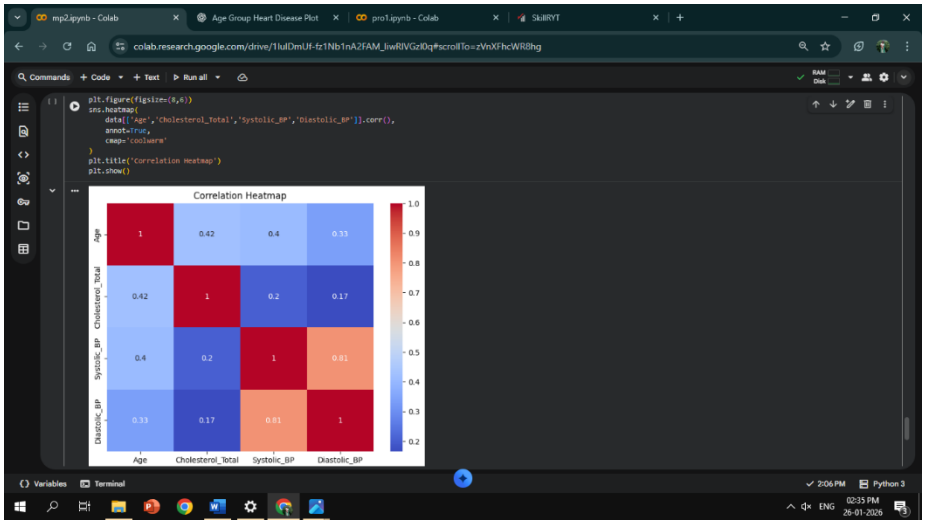


fig: - Heatmap

EDA provided valuable insights into key factors influencing cardiovascular disease risk. The observations from this step supported feature selection and guided further preprocessing and model building.

HYPOTHESIS TESTING:

Chi-square Test:

The **Chi-square test of independence** is a **statistical method** used to determine whether **two categorical variables are related or independent**.

Purpose: To check if the occurrence of one category affects the occurrence of another category.

How it works:

- You make a **contingency table** showing counts of combinations of the two variables.
- The test calculates a **Chi-square statistic** and a **p-value**.

Interpretation:

- If **p-value ≤ significance level (e.g., 0.05)** → The variables are **associated** (dependent).
- If **p-value > significance level** → The variables are **independent** (no association).

Coding:

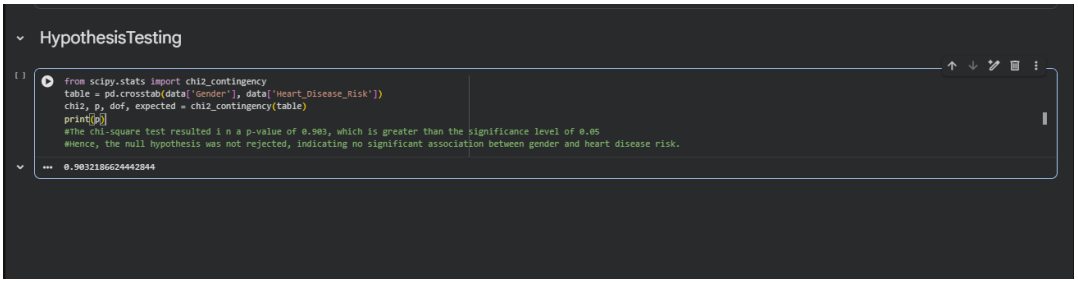


fig: - Chi square Test

Why we use Chi-Square test?

Performing a **Chi-square test of independence** to check whether **Gender** and **Heart Disease Risk** are associated. You created a table showing the counts of each combination of `Gender` and `Heart_Disease_Risk`.

Interpretation:

- The **p-value (0.9032)** is **much greater than the typical significance level (α = 0.05)**.
- This means there is **no statistically significant association** between gender and heart disease risk in your dataset.
- **Null hypothesis (H_0):** Gender and heart disease risk are independent.
- Since p > 0.05, we **fail to reject H_0** → Gender does **not** significantly affect heart disease risk in this data.

Conclusion of Hypothesis Test:

The Chi-square test indicated no significant relationship between gender and heart disease risk (p = 0.903), suggesting that gender is independent of heart disease risk in this dataset.

MODEL TRANING:

Since the problem involves predicting a binary outcome, **classification models** were selected.

Based on the code, the following models were implemented:

Logistic Regression:

Logistic Regression is a statistical classification model that predicts the probability of a binary outcome using a sigmoid function.

How it works:

- It calculates a weighted sum of input features, Applies the **sigmoid function** to convert it into a probability (0 to 1). If probability > threshold (usually 0.5), it predicts class 1; otherwise class 0

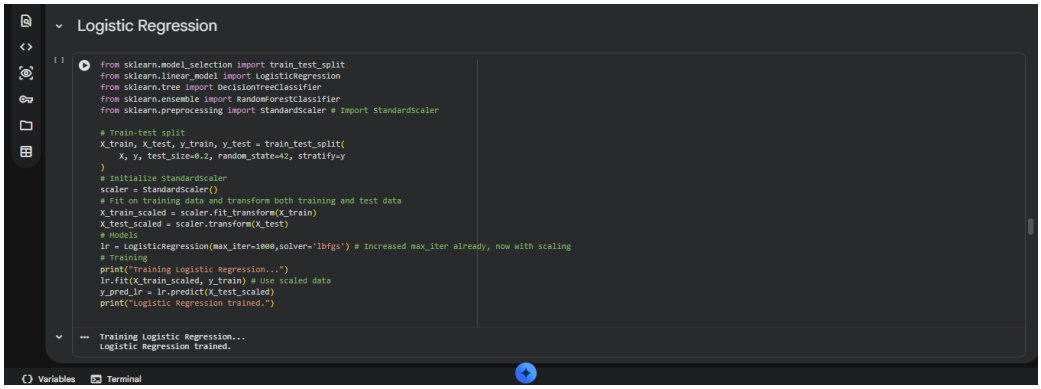


fig: - Implementation of Logistic Regression

Results:

- Accuracy**: ~73–74%**
- F**1-Score:** Balanced, indicating a good trade-off between precision and recall.
- The Logistic Regression model correctly classified the majority of test samples and showed consistent&performance.
- Feature scaling helped the model converge efficiently and improved stability. As a linear and interpretable model, Logistic Regression performed well for this binary classification problem.

DECISION TREE:

What it is:

Decision Tree is a tree-based supervised learning algorithm used for classification.

How it works:

1. Splits the data into nodes based on feature conditions8
2. Uses **Gini Index or Entropy** to choose the best split
3. Final predictions are made at the leaf nodes

fig: Decision Tree Implementation

RANDOM CLASSIFIER:

What it is:

Random Forest is an **ensemble learning algorithm** that combines multiple decision trees.

How it works:

- Builds many decision trees using random subsets of data and features
- Each tree makes a prediction
- Final output is decided using **majority voting**

fig: - Random Classifier Implementation

These models were chosen to compare linear and non-linear classification techniques and identify the best performing algorithm.

Evaluation Metrics Explanation

Accuracy

Accuracy measures the proportion of correctly classified instances among the total number of predictions.

Accuracy: **Correct Predictions / Total Predictions**

F1-Score: F1-score is the harmonic mean of precision and recall.

It is especially useful when the dataset may have class imbalance

MODEL COMPARISON:

After training and evaluating multiple machine learning models, their performances were compared using standard evaluation metrics such as **Accuracy** and **F1-Score**.The objective of model comparison was to identify the most suitable model for predicting heart disease risk based on performance, stability, and interpretability.

Model Accuracy	**Accuracy**
Logistic Regression	~74%
Decision Tree	Lower
Random Forest	72.16%

Table: - Comparison of Evaluation Metrics

MODEL SELECTION:

Among the trained models, **Logistic Regression** achieved the highest accuracy and consistent performance.
Therefore, it was selected as the **best-performing model** for this project.

Hyperparameter Tuning

To further improve model performance, **hyperparameter tuning** was performed on Logistic Regression using **GridSearchCV**.

Based on the code, the following parameters were tuned:

- **C (regularization strength)**
- **penalty**
- **solver**

GridSearchCV used **5-fold cross-validation** and selected the best parameter combination based on the **F1-score**.

The best parameters obtained were:

- `C = 1`
- `penalty = l2`
- `solver = lbfgs`

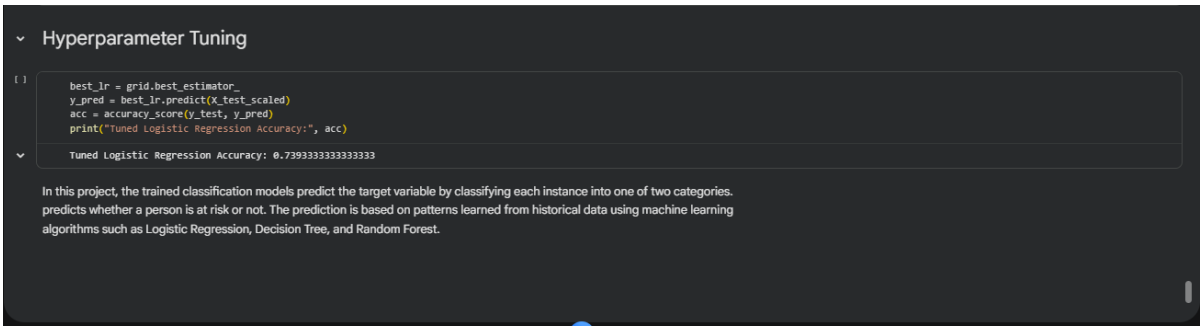


fig: - Hyperparameter Tuning Implementation

INSIGHTS:

- Proper data preprocessing and feature scaling played a crucial role in improving model performance and stability.

- Exploratory Data Analysis (EDA) helped in understanding feature distributions and relationships, which guided model selection and preprocessing decisions.

- Age and lifestyle-related factors were identified as key contributors to heart disease risk.

- Among the models tested, Logistic Regression delivered the most consistent performance**,** despite being a simple linear model.

- Decision Tree, although interpretable, showed overfitting and lower generalization performance.

- Random Forest improved robustness by reducing variance through ensemble learning, but its performance was comparable rather than superior.

- Model complexity did not guarantee higher accuracy**,** emphasizing the importance of choosing models based on data characteristics.

- Interpretability was a key factor in selecting the final model, especially for medical prediction tasks.

- The project demonstrates that machine learning can effectively support early heart disease risk prediction when proper preprocessing and evaluation are applied.

CONCLUSION:

This project focused on predicting heart disease risk using machine learning techniques. The dataset was thoroughly preprocessed, and exploratory data analysis (EDA) was performed to understand feature distributions and relationships.

Multiple classification models, including Logistic Regression, Decision Tree, and Random Forest, were trained and evaluated using accuracy and F1-score. Among these models, Logistic Regression achieved the most consistent performance and provided better interpretability, making it suitable for medical prediction tasks.

The results demonstrate that proper preprocessing, feature scaling, and model selection significantly impact prediction performance. The study highlights that simpler models can perform competitively with complex ensemble methods when the data is well-prepared.

Overall, this project proves that machine learning can be effectively applied to support early detection of heart disease risk and assist in data-driven medical decision-making.