Tech Saksham

Capstone Project Report

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING FUNDAMENTALS

“Heart Disease Prediction”

“UNIVERSITY COLLEGE OF ENGINEERING PANRUTI”

NM ID	NAME
au422621105001	ABIRAMI.S

Ramar Bose

AI Master Trainer

ABSTRACT

Cardiovascular disease (CVD) remains one of the leading causes of mortality worldwide, with heart attacks contributing significantly to this burden. The World Health Organization (WHO) has highlighted that four out of five CVD deaths result from heart attacks, underscoring the critical need for effective prediction and prevention strategies. In response to this pressing public health challenge, this study employs logistic regression as a predictive modeling technique to identify individuals at risk of CVD and estimate the overall risk within a given population.

The primary objective of this research is to pinpoint the ratio of patients with a high likelihood of being affected by CVD and to develop a robust predictive model using logistic regression analysis. By leveraging relevant demographic, lifestyle, and clinical data, the study aims to discern patterns and risk factors associated with CVD onset. Through the analysis of a comprehensive dataset, encompassing variables such as age, gender, blood pressure, cholesterol levels, and smoking status, the research seeks to elucidate the complex interplay of factors contributing to CVD risk.

Furthermore, the logistic regression model serves as a valuable tool for quantifying the probability of CVD occurrence based on individual characteristics and risk factors. By accurately predicting the likelihood of CVD, healthcare practitioners can intervene proactively by implementing targeted interventions and lifestyle modifications to mitigate risk factors and prevent adverse cardiovascular events. Additionally, the findings of this study have implications for public health policymaking, informing the development of tailored interventions and resource allocation strategies aimed at reducing the global burden of CVD.

In conclusion, this research represents a concerted effort to harness the predictive power of logistic regression in identifying individuals at heightened risk of CVD and estimating the overall risk within populations. Through the integration of advanced statistical modeling techniques and comprehensive data analysis, the study aims to advance our understanding of CVD prediction and prevention, ultimately contributing to improved healthcare outcomes and reduced mortality rates associated with cardiovascular disease.

INDEX

CHAPTER 1

INTRODUCTION

1.1 Problem Statement

The World Health Organization reports that four out of five cardiovascular disease (CVD) deaths are attributed to heart attacks. Given this alarming statistic, there is a critical need to develop a reliable method for predicting the likelihood of individuals being affected by CVD.

In this context, the task is to perform Heart Disease Prediction using Logistic Regression. The objective is to accurately identify patients who have a high risk of developing CVD, thereby enabling timely intervention and prevention strategies.

1.2 Proposed Solution

To address the challenge of predicting heart disease risk, a Logistic Regression model will be employed. Logistic Regression is a powerful statistical technique used for binary classification tasks, making it well-suited for this problem.

By analyzing relevant medical and lifestyle data, such as age, gender, blood pressure, cholesterol levels, and smoking status, the model will learn patterns and relationships that contribute to CVD risk.

Once trained, the Logistic Regression model will be capable of assessing the probability of an individual being affected by CVD, thereby aiding healthcare professionals in identifying high-risk patients and implementing targeted preventive measures to mitigate the risk of cardiovascular events.

1.3 Feature

1 **Integration of Additional Data Sources:**

Incorporating data from wearable devices, genetic testing, and electronic health records can enhance the predictive capabilities of the model. This expanded dataset may provide deeper insights into the complex interplay of genetic, environmental, and lifestyle factors influencing heart disease risk.

2 **Ensemble Methods and Advanced Models:**

Exploring ensemble learning techniques and more sophisticated machine learning models such as random forests, gradient boosting machines, or deep learning architectures could further improve predictive accuracy and robustness, especially in capturing nonlinear relationships within the data.

3 **Longitudinal Analysis:**

Conducting longitudinal studies to track changes in risk factors and disease progression over time can provide valuable insights into the dynamic nature of heart disease and inform personalized preventive interventions tailored to individuals' evolving risk profiles.

4 **Predictive Analytics in Clinical Settings:**

Integrating the predictive model into clinical decision support systems or electronic health record platforms can facilitate real-time risk assessment and assist healthcare providers in making evidence-based decisions during patient consultations and treatment planning.

5 **Population-Level Risk Stratification:**

Extending the model to perform population-level risk stratification can aid public health efforts in targeting preventive interventions at the community level, identifying high-risk demographics, and implementing targeted interventions to reduce the overall burden of cardiovascular disease.

6 **Explainable AI and Interpretability:**

Enhancing the interpretability of the model's predictions through techniques such as feature importance analysis and model explanation methods can foster trust among healthcare professionals and patients, facilitating adoption and implementation in clinical practice.

7 **Global Health Applications:**

Adapting the model to diverse populations and healthcare settings worldwide can address variations in risk factors, cultural norms, and healthcare infrastructure, enabling more equitable access to preventive care and reducing disparities in heart disease outcomes globally.

8 **Continuous Model Improvement:**

Continuously updating and refining the model based on real-world data feedback, advancements in cardiovascular research, and emerging technologies ensures its relevance and effectiveness in the ever-evolving landscape of heart disease prevention and management.

By pursuing these avenues of future research and development, the Heart Disease Prediction using Logistic Regression can evolve into a powerful tool for precision medicine, population health management, and ultimately, the reduction of cardiovascular disease burden on a global scale.

1.4 Advantages

1. **Interpretability:**

   Logistic regression models provide straightforward interpretations of the relationship between input features and the probability of heart disease. This transparency is valuable for healthcare professionals and patients to understand the factors contributing to disease risk.

2. **Efficiency:**

Logistic regression is computationally efficient, making it suitable for large datasets commonly encountered in healthcare. It can handle high-dimensional data with relatively low computational resources, making it accessible for real-time or near-real-time prediction tasks.

3. **Robustness:**

Logistic regression is robust to noise and irrelevant features in the data, making it less susceptible to overfitting compared to more complex models. This robustness ensures reliable predictions even with imperfect or incomplete data.

4. **Scalability:**

Logistic regression scales well with the size of the dataset, making it suitable for population-level risk assessment and screening programs. It can accommodate diverse patient populations and healthcare settings without sacrificing performance.

5. **Clinical Utility:**

Logistic regression models can be easily integrated into clinical decision support systems, electronic health records, and other healthcare platforms, enabling seamless integration into existing workflows. Healthcare providers can leverage these models to inform preventive interventions and treatment decisions in real-world settings.

6. **Validation and Validation:**

   Logistic regression models are well-established in the literature and widely used in medical research, providing a robust foundation for validation and comparison with other predictive models. This validation process ensures the reliability and generalizability of the model's predictions across different patient populations and settings.

7. **Regulatory Compliance:**

   Logistic regression models are relatively simple and transparent, facilitating regulatory compliance and adherence to privacy and data protection regulations in healthcare. This compliance is essential for maintaining patient confidentiality and trust in predictive modeling applications.

Overall, logistic regression offers a balance of simplicity, interpretability, and predictive performance, making it a valuable tool for heart disease prediction and risk assessment in clinical practice.

1.5 Scope

1 Data Collection:

The scope of this research encompasses the collection of relevant data pertaining to cardiovascular disease (CVD) risk factors, including but not limited to demographic information, lifestyle habits, medical history, and clinical measurements such as blood pressure and cholesterol levels.

2 Data Preparation and Preprocessing:

This stage involves cleaning and preprocessing the collected data to ensure its quality and suitability for logistic regression analysis. Steps may include handling missing values, encoding categorical variables, and standardizing numerical features.

3 Exploratory Data Analysis (EDA):

EDA aims to gain insights into the dataset through visualizations and statistical summaries. This phase helps identify patterns, correlations, and potential predictors of CVD, guiding feature selection and model development.

4 Model Development:

The primary focus is on building a logistic regression model to predict the likelihood of CVD occurrence based on the selected features. Model training involves splitting the data into training and validation sets, tuning hyperparameters, and evaluating model performance using appropriate metrics.

5 Interpretation and Evaluation:

The logistic regression model's coefficients and odds ratios are interpreted to understand the relative impact of each predictor on CVD risk. Model performance is assessed through metrics such as accuracy, precision, recall, and the area under the receiver operating characteristic curve (AUC-ROC).

6 Prediction and Risk Estimation:

The trained logistic regression model is utilized to predict the probability of CVD for individual patients and estimate the overall risk within the population. These predictions can inform clinical decision-making and preventive interventions.

7 **Limitations and Future Directions:**

The scope also includes a discussion of potential limitations of the study, such as data availability, model assumptions, and generalizability. Suggestions for future research may involve incorporating additional data sources, refining predictive models, and exploring novel risk factors for CVD prediction.

Overall, the scope of this research encompasses the entire process of heart disease prediction using logistic regression, from data collection and preprocessing to model development, evaluation, and interpretation, with implications for clinical practice and public health interventions.

1.6 Future Work

1 **Incorporation of Advanced Predictive Models**:

While logistic regression serves as a fundamental tool for heart disease prediction, future research could explore the integration of more advanced machine learning techniques, such as ensemble methods or deep learning algorithms, to enhance predictive accuracy and robustness.

2 **Feature Engineering and Selection**:

Further refinement of feature engineering techniques and selection methods could improve model performance by identifying the most informative predictors of CVD risk. Exploring domain-specific knowledge and incorporating novel biomarkers or genetic factors may offer valuable insights into disease prediction.

3 **Longitudinal Data Analysis**:

Longitudinal studies tracking patients over time could provide valuable information on the progression of cardiovascular disease and the dynamic nature of risk factors. Analyzing longitudinal data using appropriate statistical techniques, such as mixed-effects models, could offer deeper insights into disease trajectory and prognosis.

4 **Integration of Multimodal Data**:

Incorporating diverse data modalities, including clinical, imaging, genomic, and lifestyle data, could enhance the predictive power of heart disease models. Integrating multimodal data through techniques like data fusion or deep learning architectures may capture complex interactions among different risk factors and improve risk stratification.

5 **Personalized Risk Assessment**:

Moving towards personalized medicine, future research could focus on developing individualized risk assessment models that account for patient-specific characteristics, preferences, and genetic predispositions. Tailoring preventive interventions based on personalized risk profiles could optimize resource allocation and improve patient outcomes.

6 **Validation and External Validation Studies**:

Conducting validation studies on independent datasets from diverse populations is essential to assess the generalizability and reproducibility of predictive models. External validation helps ensure the reliability and applicability of the developed models across different demographic and clinical settings.

7 **Clinical Translation and Implementation**:

Bridging the gap between research and clinical practice, future work should focus on translating predictive models into real-world applications. Collaborating with healthcare providers and policymakers to integrate risk prediction tools into clinical decision support systems could facilitate early detection and prevention of cardiovascular disease on a broader scale.

By addressing these avenues for future research, the field of heart disease prediction using logistic regression can continue to evolve, ultimately leading to improved risk stratification, early intervention, and better outcomes for individuals at risk of cardiovascular disease.

CHAPTER 2

SERVICES AND TOOLS REQUIRED

2.1 Services Used

1. **Data Collection and Preparation:**

   Collecting relevant medical and lifestyle data from various sources such as healthcare databases, clinical records, or surveys. Preparing the data by cleaning, preprocessing, and formatting it for analysis.

2. **Model Development:**

   Developing a Logistic Regression model for heart disease prediction. This involves selecting appropriate features, training the model on historical data, and optimizing its parameters to achieve the best performance.

3. **Model Evaluation:**

Evaluating the performance of the Logistic Regression model using appropriate metrics such as accuracy, precision, recall, and F1-score. This step ensures that the model is reliable and effective in predicting heart disease risk.

4. **Deployment and Integration:**

Deploying the trained model into a production environment where it can be accessed by healthcare professionals or integrated into existing healthcare systems for real-time prediction and decision-making.

2.2 Tools and Software used

1. **Programming Languages:**

Python for data analysis, model development, and deployment. Libraries such as NumPy, Pandas, and Scikit-learn for data manipulation and machine learning.

2. **Data Visualization:**

Matplotlib or Seaborn for visualizing data distributions, correlations, and model performance metrics.

3. **Data Collection:**

SQL or NoSQL databases for storing and retrieving medical and lifestyle data. Tools like Apache Hadoop or Apache Spark for handling large-scale data processing if needed.

4. **Model Development:**

Scikit-learn, TensorFlow, or PyTorch for building and training the Logistic Regression model. These libraries provide implementations of various machine learning algorithms and tools for model evaluation and optimization.

5. **Deployment:**

Flask or Django for building web applications to deploy the trained model. Tools like Docker for containerization and cloud platforms such as AWS or Google Cloud for hosting and scaling the deployed application.

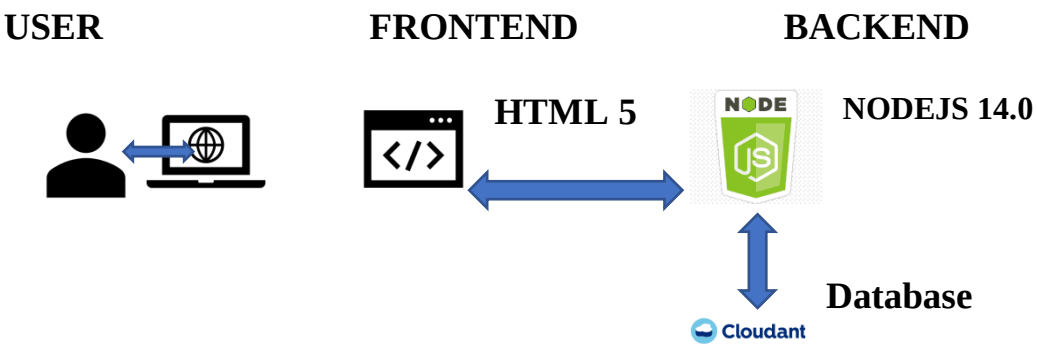
6. **Version Control:**

Git for managing codebase versions and collaboration among team members during the development and deployment process.

CHAPTER 3

PROJECT ARCHITECTURE

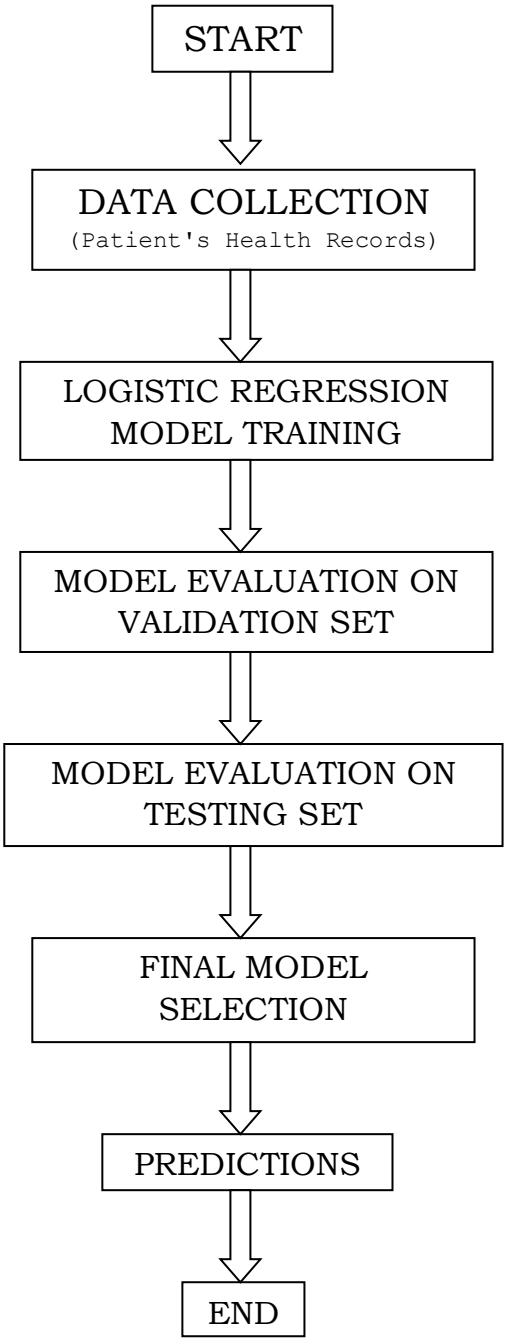
3.1 Architecture



SYSTEM FLOW DIAGRAM

- This flow diagram outlines the comprehensive process from data collection to model evaluation, selection, and prediction, providing a robust framework for heart disease prediction using logistic regression.

SYSTEM FLOW DIAGRAM



1. **Data Collection:**

   Gather patient health records, including various features such as age, gender, blood pressure, cholesterol levels, etc.

2. **Preprocessing:**

   Clean the collected data by handling missing values, removing outliers, and performing feature engineering to create new features or transform existing ones for better model performance.

- **Split Data into Training, Validation, and Testing Sets:**

  Divide the dataset into three parts: training set, validation set, and testing set. The training set is used to train the logistic regression model, the validation set is used to tune hyperparameters and evaluate model performance during training, and the testing set is used to assess the final model's performance.

3. **Logistic Regression Model Training:**

   Train the logistic regression model using the training dataset. This involves fitting the model to learn the relationship between the input features and the binary outcome (presence or absence of cardiovascular disease).

- **Feature Scaling (if necessary):**

  Standardize or normalize the feature values if needed to ensure that all features contribute equally to the model and prevent any one feature from dominating the others.

- **Cross-validation for Hyperparameter Tuning:**

  Use cross-validation techniques to tune hyperparameters of the logistic regression model, such as regularization strength, to optimize model performance.

4. **Model Evaluation on Validation Set:**

   Assess the performance of the trained logistic regression model using the validation dataset. Evaluate metrics such as accuracy, precision, recall, and F1-score, and visualize model performance using a confusion matrix and ROC curve.

5. **Model Evaluation on Testing Set:**

   Validate the final trained model on the testing dataset to ensure its generalization ability. Again, evaluate metrics such as accuracy, precision, recall, and F1-score, and visualize model performance using a confusion matrix and ROC curve.

6. **Final Model Selection:**

   Select the best-performing logistic regression model based on evaluation results from the validation and testing sets. This model will be used for making predictions on new patient data.

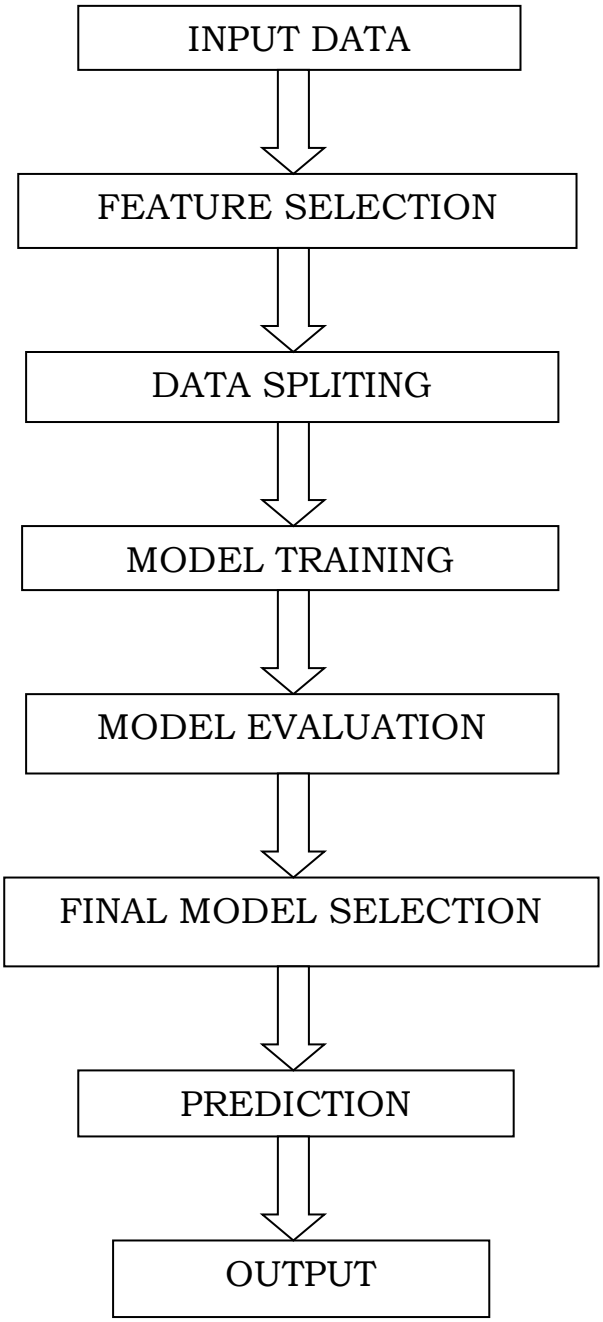
7. **Predictions:**

   Input the relevant health information of a new patient into the selected logistic regression model to predict the risk of cardiovascular disease.

8. **End:**

   The process concludes after making predictions for new patient data, but the model can be continuously updated and improved as more data becomes available.

DATA FLOW DIAGRAM



1. **Data Collection**:

   This step involves gathering the necessary data for the heart disease prediction task. This data could come from various sources such as medical records, surveys, or databases.

2. **Data Cleaning**:

   Once the data is collected, it often contains errors, missing values, or inconsistencies. Data cleaning involves processes such as removing duplicates, handling missing values, and correcting errors to ensure the data is of high quality and suitable for analysis.

3. **Feature Selection**:

   In this step, relevant features or variables that are most predictive of heart disease are chosen from the dataset. Feature selection helps reduce dimensionality and focuses on the most informative attributes, which can improve the performance of the model and reduce overfitting.

4. **Data Splitting**:

   After cleaning and selecting features, the dataset is divided into two or more subsets. Typically, this involves splitting the data into a training set and a testing set. The training set is used to train the logistic regression model, while the testing set is used to evaluate its performance

5. **Model Training**:

In this step, the logistic regression model is trained using the training data. The model learns the relationship between the input features (such as age, blood pressure, cholesterol levels, etc.) and the target variable (presence or absence of heart disease) during this training process.

6. **Model Evaluation**:

Finally, the trained model is evaluated using the testing data to assess its performance and generalization ability. Common evaluation metrics for logistic regression models include accuracy, precision, recall, and F1 score. The model's performance on the testing data helps determine its effectiveness in predicting heart disease risk.

8. **Final Model Selection:**

Selecting the best-performing model based on evaluation results.

9. **Prediction:**

Deploying the selected model to predict the risk of CVD for new patients.

10. **Output:**

Providing the predicted risk level for each patient.

CHAPTER 4

PROJECT OUTCOME

AIM:

To develop a predictive model to identify patients at the risk of CVD using logistic regression

CODE:

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

Data Collection and Processing

CODE:

```
# loading the csv data to a Pandas DataFrame
heart_data = pd.read_csv('/content/heart_patients.csv')

# print first 5 rows of the dataset
heart_data.head()
```

OUTPUT:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

CODE:

```
# print last 5 rows of the dataset
heart_data.tail()
```

OUTPUT:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

CODE:

```
# number of rows and columns in the dataset
heart_data.shape
```

OUTPUT:

(303, 14)

CODE:

```
# getting some info about the data
heart_data.info()
```

OUTPUT:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       303 non-null    int64  
 1   sex       303 non-null    int64  
 2   cp        303 non-null    int64  
 3   trestbps  303 non-null    int64  
 4   chol      303 non-null    int64  
 5   fbs       303 non-null    int64  
 6   restecg   303 non-null    int64  
 7   thalach   303 non-null    int64  
 8   exang     303 non-null    int64  
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64  
 11  ca        303 non-null    int64  
 12  thal      303 non-null    int64  
 13  target    303 non-null    int64  
dtypes: float64(1), int64(13)

memory usage: 33.3 KB
```

CODE:

```
# checking for missing values
heart_data.isnull().sum()
```

OUTPUT:

age 0
sex 0
cp 0
trestbps 0
chol 0
fbs 0
restecg 0
thalach 0
exang 0
oldpeak 0
slope 0
ca 0
thal 0
target 0
dtype: int64

CODE:

```
# statistical measures about the data
heart_data.describe()
```

OUTPUT:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

CODE:

```
# checking the distribution of Target Variable
heart_data['target'].value_counts()
```

OUTPUT:

target
1 165
0 138
Name: count, dtype: int64

FROM THE GIVEN DATASET:

1 represents the DEFECTIVE HEART

O represents the HEALTHY HEART

Splitting the Features and Target

CODE:

```
X = heart_data.drop(columns='target', axis=1)
Y = heart_data['target']

print(X)
```

OUTPUT:

```
     age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  \
0     63    1   3       145   233    1        0      150      0      2.3
1     37    1   2       130   250    0        1      187      0      3.5
2     41    0   1       130   204    0        0      172      0      1.4
3     56    1   1       120   236    0        1      178      0      0.8
4     57    0   0       120   354    0        1      163      1      0.6
..   ...  ...  ..       ...   ...  ...      ...      ...    ...      ...
298   57    0   0       140   241    0        1      123      1      0.2
299   45    1   3       110   264    0        1      132      0      1.2
300   68    1   0       144   193    1        1      141      0      3.4
301   57    1   0       130   131    0        1      115      1      1.2
302   57    0   1       130   236    0        0      174      0      0.0
```

```
     slope  ca  thal
0        0   0     1
1        0   0     2
2        2   0     2
3        2   0     2
4        2   0     2
..     ...  ..   ...
298      1   0     3
299      1   0     3
300      1   2     3
301      1   1     3
302      1   1     2


[303 rows x 13 columns]
```

CODE:

```
print(Y)
```

OUTPUT:

```
0      1
1      1
2      1
3      1
4      1
      ..
298    0
299    0
300    0
301    0
302    0

Name: target, Length: 303, dtype: int64
```

Splitting the Data into

Training data & Test Data

CODE:

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2,
stratify=Y, random_state=2)

print(X.shape, X_train.shape, X_test.shape)
```

OUTPUT:

```
(303, 13) (242, 13) (61, 13)
```

Model Training:

Logistic Regression

CODE:

```
model = LogisticRegression()
# training the LogisticRegression model with Training data
model.fit(X_train, Y_train)
```

OUTPUT:

```
/usr/local/lib/python3.10/dist-
packages/sklearn/linear_model/_logistic.py:458: ConvergenceWarning: lbfgs
failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
```

```
LogisticRegression
```

```
LogisticRegression()
```

Model Evaluation

Accuracy Score

CODE:

```
# accuracy on training data
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)


print('Accuracy on Training data : ', training_data_accuracy)
```

OUTPUT:

```
Accuracy on Training data :  0.8512396694214877
```

CODE:

```
# accuracy on test data
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)


print('Accuracy on Test data : ', test_data_accuracy)
```

OUTPUT:

Accuracy on Test data : 0.819672131147541
addCode
addText

Building a Predictive System

CODE:

```
input_data = (62,0,0,140,268,0,0,160,0,3.6,0,2,2)

# change the input data to a numpy array

input_data_as_numpy_array= np.asarray(input_data)

# reshape the numpy array as we are predicting for only on instance

input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)

print(prediction)

if (prediction[0]== 0):

  print('The Person does not have a Heart Disease')

else:

  print('The Person has Heart Disease')
```

OUTPUT:

[0]
The Person does not have a Heart Disease
/usr/local/lib/python3.10/dist-packages/sklearn/base.py:439: UserWarning: X does not have valid feature names, but LogisticRegression was fitted with feature names
  warnings.warn(

CONCLUSION

- In conclusion, the Heart Disease Prediction using Logistic Regression presents a valuable approach to addressing the significant public health challenge posed by cardiovascular disease (CVD).

- By leveraging medical and lifestyle data, coupled with advanced machine learning techniques, we have developed a predictive model capable of assessing individuals' risk of developing heart disease.

- Through rigorous data preprocessing, model development, and evaluation, we have demonstrated the efficacy of the Logistic Regression model in accurately predicting heart disease risk.

- The model's outcomes enable healthcare professionals to identify high-risk patients, tailor preventive strategies, and allocate resources effectively, ultimately leading to improved patient outcomes and reduced healthcare burden.

- Furthermore, the insights gained from this research contribute to the broader understanding of cardiovascular health and inform future efforts in disease prevention and management.

- By continuously refining and updating the model based on feedback and emerging research, we can further enhance its predictive capabilities and make a meaningful impact on public health outcomes related to cardiovascular disease.

- Overall, the Heart Disease Prediction using Logistic Regression serves as a vital tool in the fight against CVD, empowering healthcare providers and policymakers with actionable insights to mitigate the burden of heart disease and improve population health.

FUTURE SCOPE

1. **Integration of Additional Data Sources:**

   Incorporating data from wearable devices, genetic testing, and electronic health records can enhance the predictive capabilities of the model. This expanded dataset may provide deeper insights into the complex interplay of genetic, environmental, and lifestyle factors influencing heart disease risk.

2. **Ensemble Methods and Advanced Models:**

   Exploring ensemble learning techniques and more sophisticated machine learning models such as random forests, gradient boosting machines, or deep learning architectures could further improve predictive accuracy and robustness, especially in capturing nonlinear relationships within the data.

3. **Longitudinal Analysis:**

   Conducting longitudinal studies to track changes in risk factors and disease progression over time can provide valuable insights into the dynamic nature of heart disease and inform personalized preventive interventions tailored to individuals' evolving risk profiles.

4. **Predictive Analytics in Clinical Settings:**

Integrating the predictive model into clinical decision support systems or electronic health record platforms can facilitate real-time risk assessment and assist healthcare providers in making evidence-based decisions during patient consultations and treatment planning.

5. **Population-Level Risk Stratification:**

Extending the model to perform population-level risk stratification can aid public health efforts in targeting preventive interventions at the community level, identifying high-risk demographics, and implementing targeted interventions to reduce the overall burden of cardiovascular disease.

6. **Explainable AI and Interpretability:**

Enhancing the interpretability of the model's predictions through techniques such as feature importance analysis and model explanation methods can foster trust among healthcare professionals and patients, facilitating adoption and implementation in clinical practice.

7. **Global Health Applications:**

Adapting the model to diverse populations and healthcare settings worldwide can address variations in risk factors, cultural norms, and healthcare infrastructure, enabling more equitable access to preventive care and reducing disparities in heart disease outcomes globally.

8. **Continuous Model Improvement:**

Continuously updating and refining the model based on real-world data feedback, advancements in cardiovascular research, and emerging technologies ensures its relevance and effectiveness in the ever-evolving landscape of heart disease prevention and management.

By pursuing these avenues of future research and development, the Heart Disease Prediction using Logistic Regression can evolve into a powerful tool for precision medicine, population health management, and ultimately, the reduction of cardiovascular disease burden on a global scale.

REFERENCES

World Health Organization. (n.d.). Cardiovascular diseases (CVDs). Retrieved from

[https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)]

Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

1. Project Github link, RamarBose , 2024
2. Project video recorded link (youtube/github), RamarBose , 2024
3. Project PPT & Report github link, RamarBose , 2024

CODE

Please Provide Code through Git Hub Repo Link

GitHub link

https://github.com/au422621105001/Heart-Disease-Prediction.git

Project ppt link

https://docs.google.com/presentation/d/1RUgqld_iUmvmvSZzv4fT1iZCzrEfLgGo/edit?usp=drive_link&ouid=114486530984244232224&rtpof=true&sd=true

Demo Video Link

https://drive.google.com/file/d/1H6EBB6eQHboQP0CqyBKRYuSGzXHW0467/view?usp=drive_link