

**Tech Saksham**

**Capstone Project Report**

**ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING FUNDAMENTALS**

**“Heart Disease Prediction using Logistic Regression”**

**“UNIVERSITY COLLEGE OF ENGINEERING (BIT CAMPUS) TIRUCHIRAPALLI”**

|  |  |
| --- | --- |
| **NM ID** | **NAME** |
| au810021214003 | AKSHAYA S |

|  |  |
| --- | --- |
|  |  |
|  | Trainer Name  RAMAR BOSE |
|  | Sr. AIMaster Trainer |

**ABSTRACT**

This Logistic Regression Model of Machine learning aims to develop a predictive model to identify individuals at risk of heart disease. The major goal of this study is to identify the proportion of patients who are at high risk of developing CVD. Using a dataset of pertinent medical and demographic variables, we will use Logistic Regression, a sophisticated statistical technique well-suited for binary classification tasks, to estimate the likelihood of an individual acquiring heart disease. The project's key processes include data preparation, feature selection, model training, and evaluation. We will rigorously preprocess the dataset, removing missing values, encoding category variables, and scaling numerical features as needed. The developed model has the potential to help healthcare providers with early intervention and tailored risk assessment, ultimately lowering CVD-related morbidity and mortality rates.

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Table of Contents** | **Page No.** |
| 1 | Chapter 1: Introduction | 1 |
| 2 | Chapter 2: Services and Tools Required | 3 |
| 3 | Chapter 3: Project Architecture | 4 |
| 4 | Chapter 4: Modeling and Project Outcome | 6 |
| 5 | Chapter 5: Project Result | 12 |
| 6 | Conclusion | 13 |
| 7 | Future Scope | 14 |
| 8 | References | 15 |
| 9 | Links | 16 |

**CHAPTER 1**

**INTRODUCTION**

* 1. **Problem Statement**

The goal of this study is to estimate the total risk of heart disease using Logistic Regression and Machine Learning. Cardiovascular diseases (CVD) are a serious global health risk, with heart attacks being the leading cause of death. According to the World Health Organization (WHO), heart attacks account for about four out of every five CVD-related deaths. Despite advances in medical science, timely identification of vulnerable individuals remains a difficulty. Identifying those at risk of CVD is critical for implementing preventative strategies and giving timely therapies

* 1. **Proposed Solution**

The suggested Logistic Regression in Google's collaboration with Heart Disease Prediction includes data processing, feature selection, and model construction. Using the Logistic Regression model of MI, it is possible to assess an individual's risk of developing cardiovascular disease. This model can help healthcare practitioners identify high-risk patients and implement targeted therapies and lifestyle changes to reduce the risk of CVD-related consequences. The key processes include data preparation, feature selection, model training, and evaluation. Feature selection approaches such as correlation analysis and domain knowledge expertise will help us discover the most useful predictors in our model. Model performance will be evaluated using criteria such as accuracy and precision.

* 1. **Feature**

1. **Data Preprocessing:** The dataset undergoes preprocessing steps such as handling missing values, encoding categorical variables, and scaling numerical features.
2. **Feature Selection:** Relevant features are selected based on their significance in predicting CVD risk, using techniques such as correlation analysis and feature importance ranking**.**
3. **Model Development:** A Logistic Regression model is trained using the preprocessed dataset. The model learns to classify individuals into two classes: those at risk of developing CVD and those not at risk.
   1. **Advantages**

* One of the primary benefits of utilizing Logistic Regression for heart disease prediction is its interpretability. Unlike more complicated machine learning models, Logistic Regression explains the association between input data and the chance of heart disease.
* Logistic Regression is a computationally efficient and simple algorithm, making it ideal for situations with limited computational resources or rapid deployment, like clinical settings.
* Logistic Regression is a model that can effectively use diverse patient data, including demographics, medical history, and diagnostic test results, to improve prediction accuracy.
* Logistic Regression offers probabilistic predictions for heart disease risk stratification, enabling healthcare providers to prioritize interventions for patients with the highest predicted risk.
  1. **Scope**

The future of Logistic Regression-based heart disease prediction appears promising. When additional risk factors and powerful machine learning approaches are combined, advanced prediction models are likely to develop, resulting in better accuracy and reliability in identifying people at risk of cardiovascular disease (CVD). Furthermore, the concept of customized medicine is set to change heart disease management. Logistic regression models assist insurers in estimating heart disease risk profiles by assessing demographic, lifestyle, and medical history data, allowing for appropriate premium pricing and tailored wellness initiatives. Telemedicine and remote patient monitoring technologies allow Logistic Regression models to estimate heart disease risk, notify healthcare professionals early, permit prompt interventions, and reduce hospital readmissions.

**CHAPTER 2**

**SERVICES AND TOOLS REQUIRED**

**2.1 Services Used**

**Data Collection and Storage Services**: Access and use healthcare databases like MIMIC-III, NHANES, or the Framingham Heart Study dataset, which contain anonymised patient data related to heart disease or may also use cloud storage platforms like Github or Kaggle.

**Data Processing Services**: Python packages like Pandas, NumPy, and scikit-learn are used for data cleaning, preprocessing, and feature engineering and Matplotlib with Seaborn to visualize data distributions, correlations, and trends.

**Machine Learning Services**: Machine learning frameworks such as google collab, jupyter, scikit-learn, or PyTorch to implement and train Logistic Regression models on a dataset.

**2.2 Tools and Software used**

* **Data collection and storage:**

**Databases:** Github and Kaggle are used and the heart disease prediction dataset are downloaded.

* **Data Processing:**

Python libraries like Pandas, NumPy, scikit-learn are used and Matplotlib for

visualization used

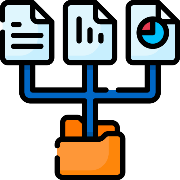
* **Model Evaluation and interpretation**:

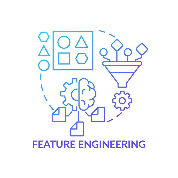
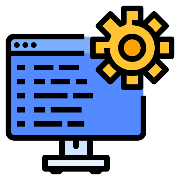
Google collab software is used to run the program ad the results are predicted.

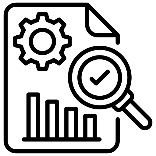
**CHAPTER 3**

**PROJECT ARCHITECTURE**

**3.1 Architecture**

 **Data Collection > Data Preprocessing >**

** Feature Engineering >** **Model Development >**

** Model Evaluation >**

Here’s a high-level architecture for the project:

**Data Collection**: Collect a broad dataset that includes demographic information, lifestyle factors (such as smoking, diet, and exercise habits), clinical parameters (such as blood pressure, cholesterol levels, and CVD family history), and outcomes (CVD presence or absence).   
**Data preprocessing**: Clean up the dataset by removing missing values, encoding category variables, and normalizing numerical characteristics.   
Exploratory data analysis (EDA): Examine the dataset for trends, correlations, and outliers that could influence CVD risk.   
**Model Development**: Use logistic regression to create a predictive model that assesses the likelihood of CVD occurrence depending on input features.   
**Model Evaluation**: Evaluate the performance of the logistic regression model using appropriate evaluation measures, such as accuracy and precision.

**Interpretation and Insights**: Analyze the logistic regression model coefficients to determine the impact of various risk factors on CVD risk and provide practical conclusions for preventive healthcare.

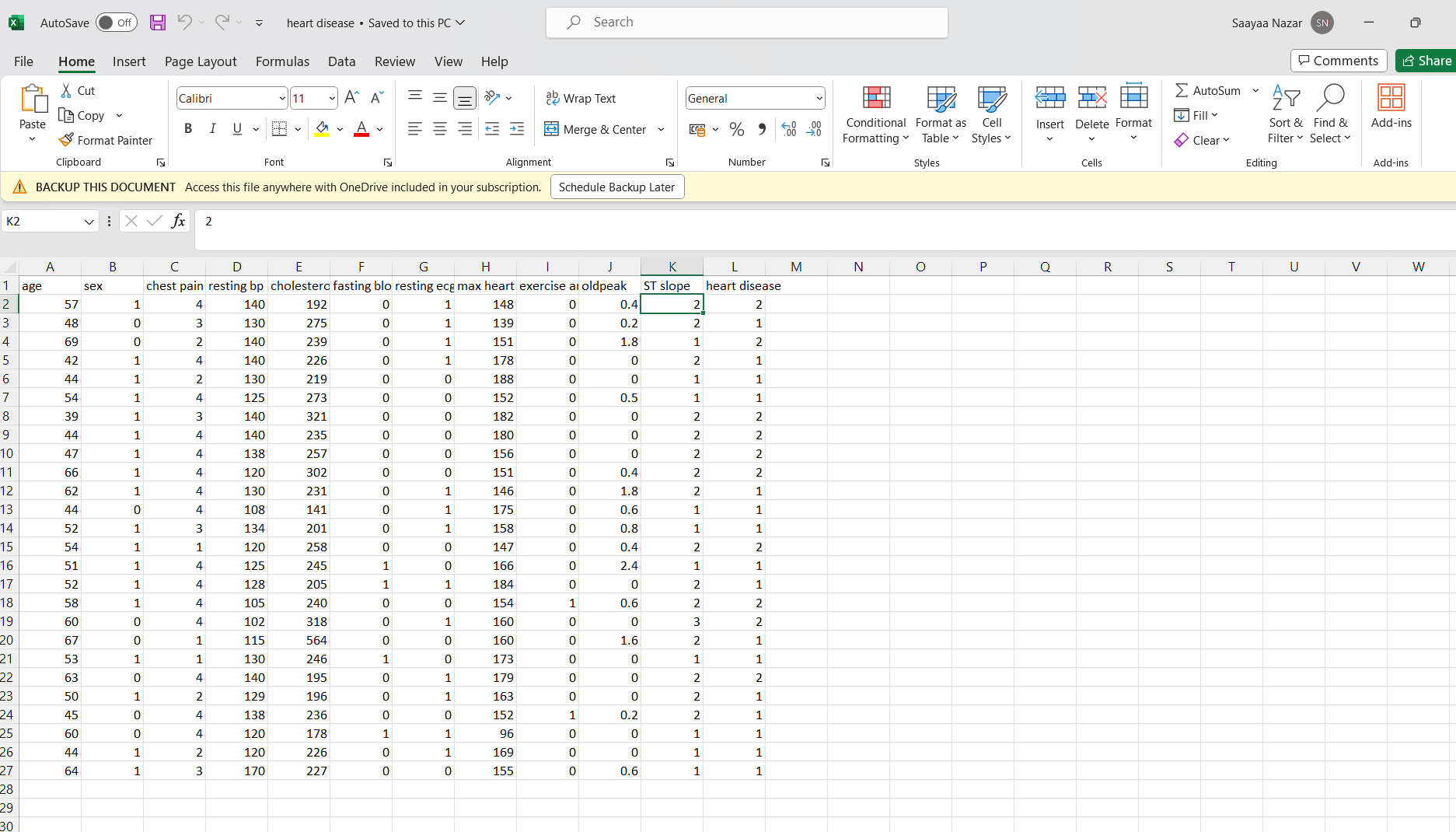
This architecture provides vital insights into Logistic Regression's prediction skills in identifying individuals at risk of heart disease. Furthermore, the created model may assist healthcare providers in early intervention and tailored risk assessment, ultimately contributing to a reduction in CVD-related morbidity and mortality rates.

**CHAPTER 4**

**MODELING AND PROJECT OUTCOME**

**Datasets**

1. The datasets are retrieved from the GitHub platform and analyzed

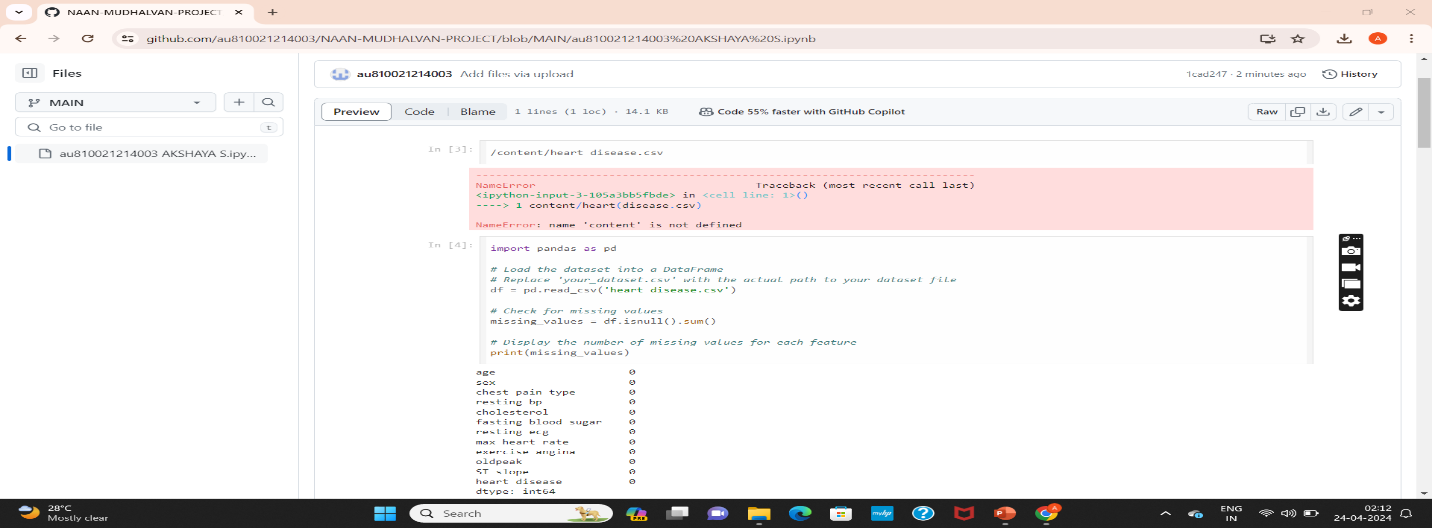


Google collab software is used to run the program and predict the heart disease in the people community

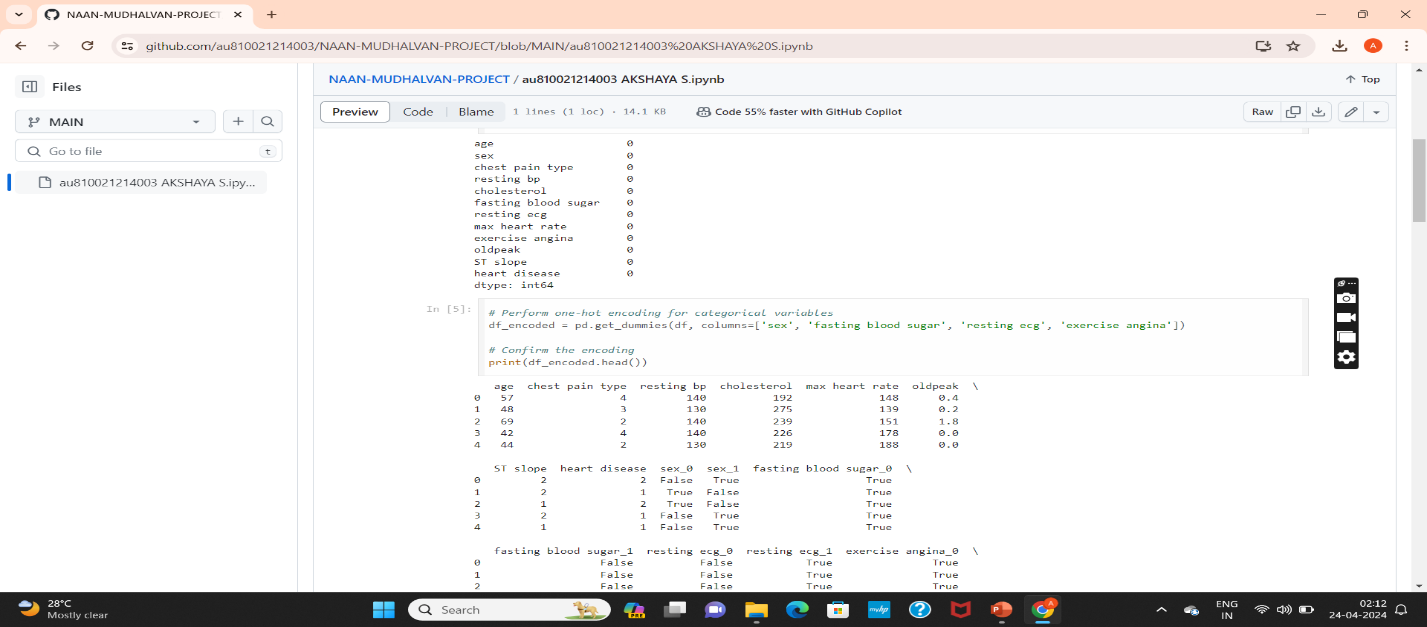
Code:

* Data Collection
* Data Preprocessing

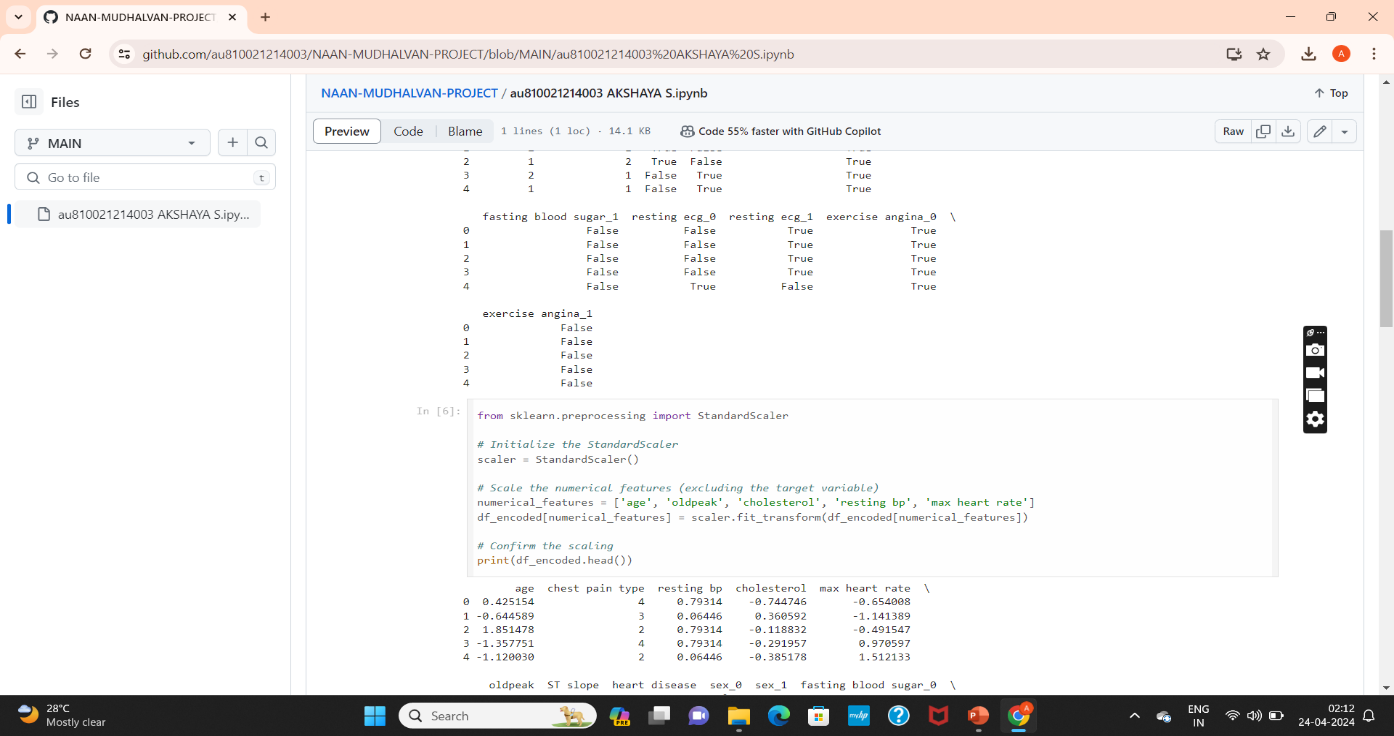
1. Import packages (Numpy, Pandas) for processing the dataset from GitHub.



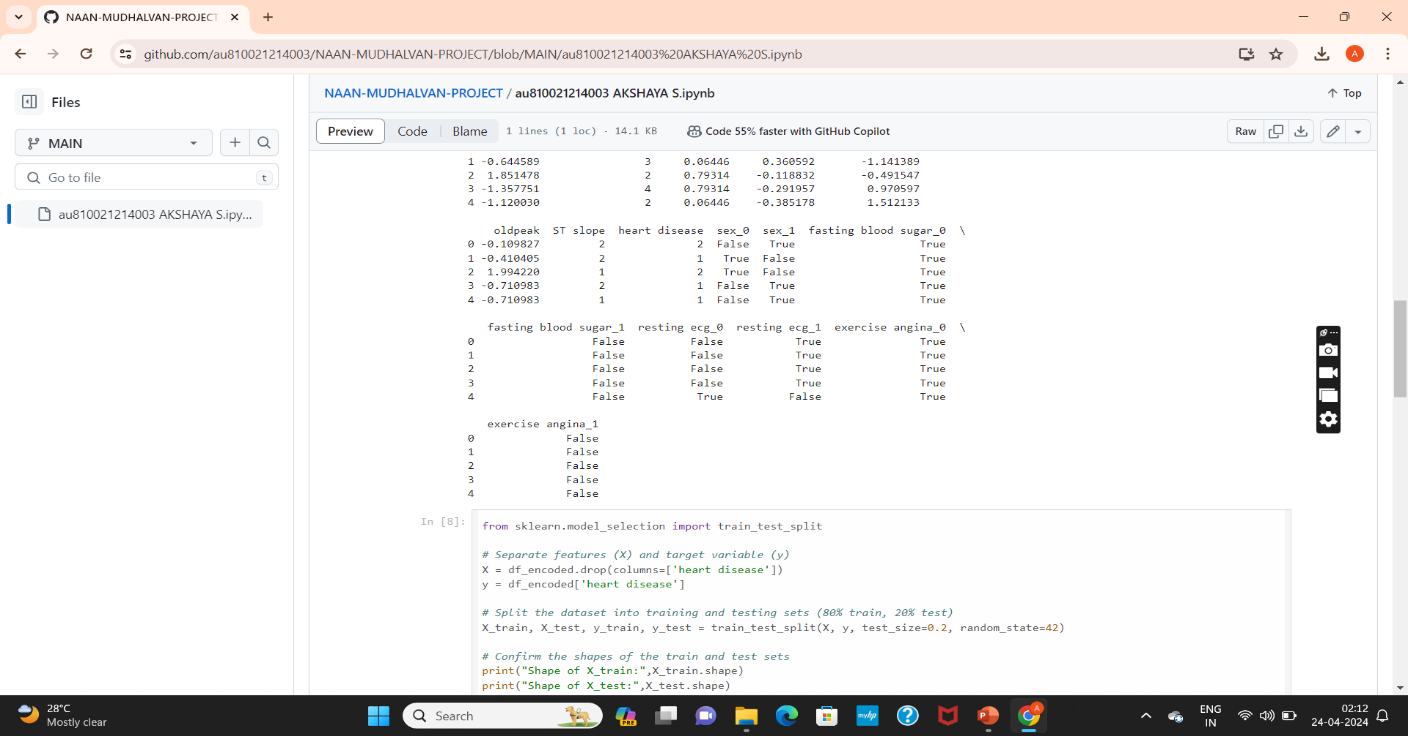
**Output:**

****

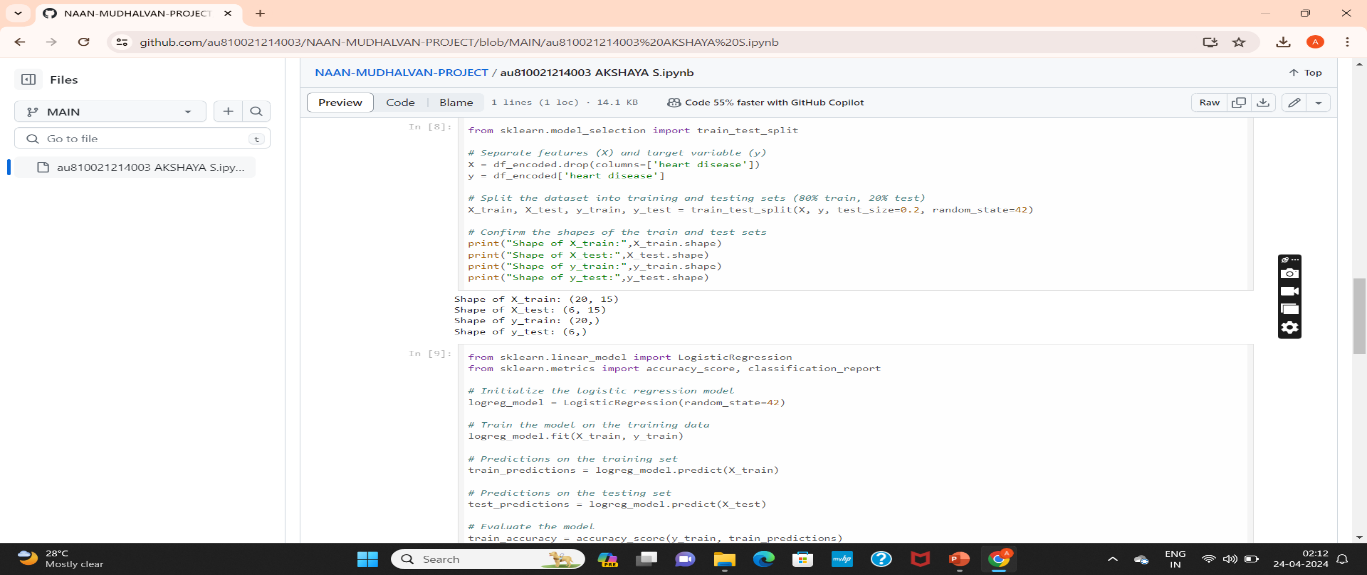
1. This code snippet performs one-hot encoding for categorical variables in the DataFrame df, where can visually inspect how the categorical variables have been transformed into binary features.



1. To standardize the numerical features in the DataFrame

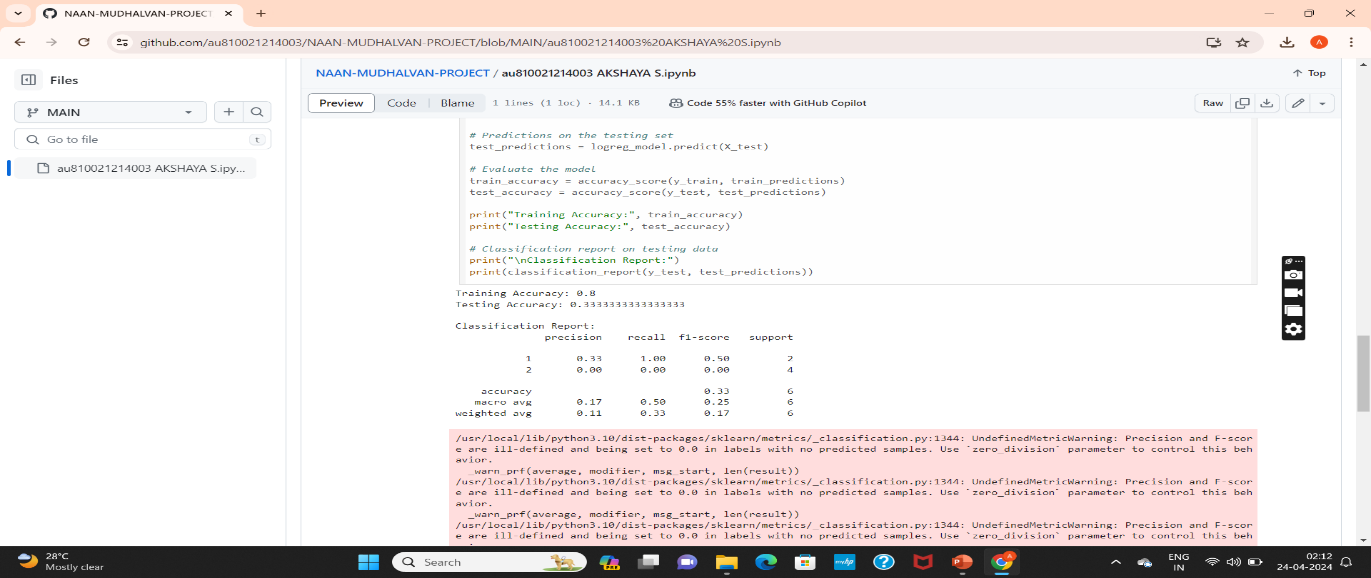


1. Prepares the dataset for model training and evaluation by splitting it into separate sets for training and testing, allowing for unbiased assessment of the model's performance on unseen data.

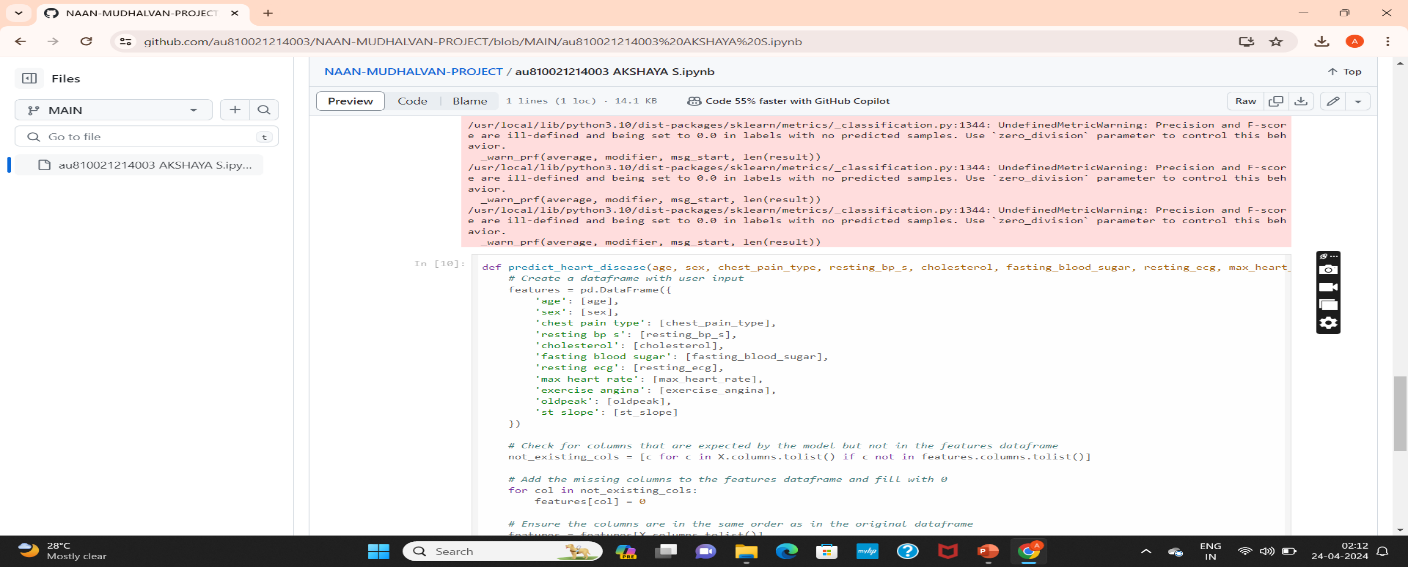


1. A Logistic Regression model on the training data, making predictions on both the training and testing sets, and evaluating the model's performance using accuracy score and classification report

Code:

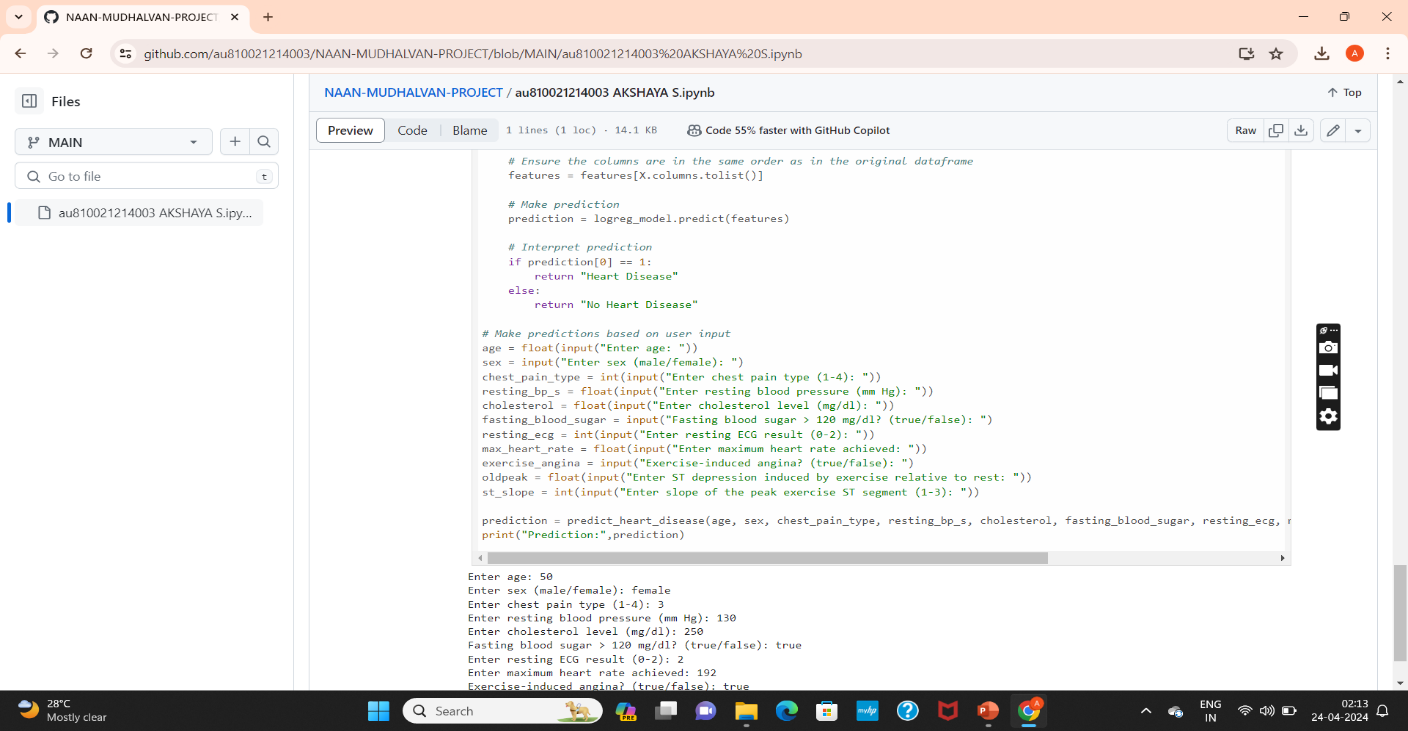


Output:

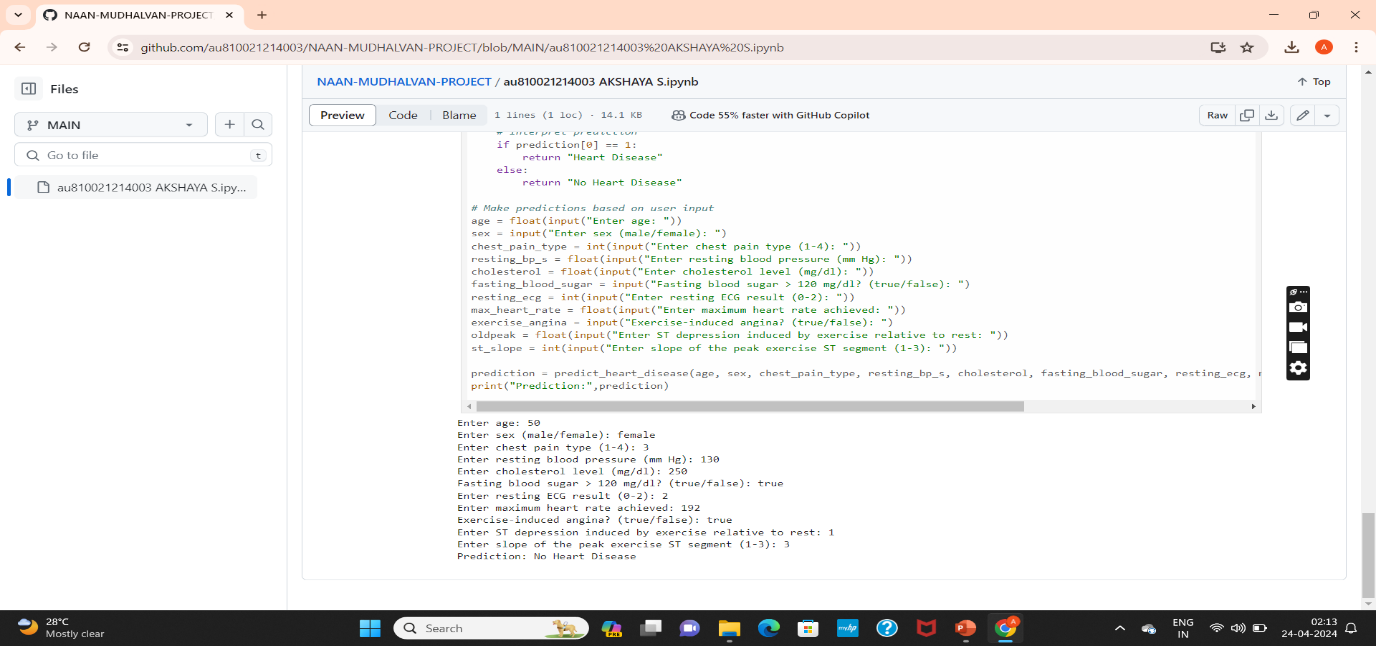


1. This takes user input for various attributes related to heart health, creates a DataFrame with the input, and uses a pre-trained Logistic Regression model to predict whether the user is likely to have heart disease or not.

Code:



Output:



**CHAPTER 5**

**Project result**

The Logistic Regression model had a training and the model achieved a testing accuracy on the testing set. The classification report contains precise performance indicators for the model, such as precision, recall, F1-score, and support for each class. The Logistic Regression model created in this project shows potential in predicting heart disease using demographic and clinical data. More modification and optimization of the model, as well as validation on additional datasets, may be required to improve its performance and generalizability. The achieved accuracy scores show that the model does a reasonable job of predicting heart disease based on the features provided. The classification report reveals additional insights into the model's performance, such as its ability to correctly classify positive and negative instances, and its balance between precision and recall.

**CONCLUSION**

The application of Logistic Regression to predict heart disease risk provides useful insights into identifying those who are more likely to develop cardiovascular disease (CVD) and maybe avoid negative health outcomes. Logistic Regression is used to successfully model the probability of CVD occurrence based on numerous risk variables, allowing for early intervention and focused healthcare interventions. The findings highlight the importance of proactive healthcare activities, such as lifestyle adjustments, regular screenings, and individualized therapies, to reduce CVD risk factors and improve patient outcomes. The Logistic Regression model developed in this project shows promise in predicting heart disease based on demographic and clinical factors. Further refinement and optimization of the model, as well as validation on additional datasets, may be warranted to enhance its performance and generalizability.

**FUTURE SCOPE**

The future of Logistic Regression-based heart disease prediction appears promising. When additional risk factors and powerful machine learning approaches are combined, advanced prediction models are likely to develop, resulting in better accuracy and reliability in identifying people at risk of cardiovascular disease (CVD). Furthermore, the concept of customized medicine is set to change heart disease management. Investigating personalized medicine approaches that tailor predictive models to individual patient characteristics, such as genetic predispositions, lifestyle factors, and medical history, could result in more accurate and targeted risk assessments, allowing for personalized prevention and treatment strategies.

**REFERENCES**

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

# **GIT Hub Link of Project Code:**

<https://github.com/au810021214003/NAAN-MUDHALVAN-PROJECT>

YOUTUBE LINK:

https://youtu.be/jMDQXQm1TDE