**** CANTILEVER AIML PROTERNSHIP 2025

**ABSTRACT**

Heart disease remains one of the leading causes of death worldwide. Early detection can significantly improve patient outcomes and reduce mortality rates. This project presents a machine learning-based solution that predicts the likelihood of a person having heart disease using various clinical parameters such as age, blood pressure, cholesterol levels, and more. The model was trained using a publicly available heart disease dataset and deployed through an interactive Streamlit web application. The application enables users to input their medical details and receive instant feedback, making it a useful tool for awareness and preliminary screening.

**INTRODUCTION**

Heart disease is a major global health concern and one of the leading causes of death. Early prediction can help in timely treatment and potentially save lives. With the help of machine learning, we can build models that predict the risk of heart disease based on medical data.

This project uses a supervised learning model trained on a publicly available dataset to predict the likelihood of heart disease in a person. The model considers various health parameters such as age, gender, blood pressure, cholesterol, chest pain type, and more.

The trained model is deployed using **Streamlit**, allowing users to input their medical data and receive predictions in a user-friendly web application. This tool aims to raise awareness and assist in early detection, not to replace professional medical advice.

**OBJECTIVES**

The primary objective of this project is to develop an intelligent system that can predict whether an individual is at risk of heart disease based on various medical parameters. By leveraging machine learning algorithms, the system aims to assist in early diagnosis and provide a quick, accessible screening tool for users.

**Goals**:

* To build and train a machine learning model using labeled heart disease data.
* To create an interactive web interface using Streamlit for easy user interaction.
* To provide immediate prediction results based on user inputs.
* To help raise awareness about heart health by encouraging users to check their risk levels.
* To demonstrate the practical application of AI/ML in real-world healthcare scenarios.

**TOOLS AND TECHNOLOGIES USED**

**Languages and Frameworks**:

* Python: The core programming language used for model development and web app creation.
* Streamlit: A Python-based framework used to build the interactive web app interface.

**Libraries and Modules**:

* sk-learn: Used to train and evaluate the machine learning model (e.g., logistic regression, random forest).
* pickle: Used for saving and loading the trained model (model\_heart.pkl).
* pandas and numpy *(used during model training)*: For handling datasets and performing numerical computations (though not in the app, used offline for preprocessing).

**Model deployement**:

* The model is stored in a .pkl file and integrated into the Streamlit app using Python’s pickle module.

**User Interface Features**:

* **st.radio()**: For selecting categorical inputs like gender, chest pain type, thalassemia, etc.
* **st.text\_input()**: For numerical inputs like age, cholesterol, and blood pressure.
* **st.success(), st.error(), st.balloons(), st.snow()**: For visually engaging prediction output.

**DATASET DESCRIPTION**

This project uses the **Heart Disease Dataset**, originally from the UCI Machine Learning Repository and hosted on Kaggle. It is designed to predict the presence of heart disease in patients based on medical attributes.

**Context**:

The original dataset combines data from four different sources: **Cleveland, Hungary, Switzerland, and Long Beach V**. Although the full dataset contains 76 attributes, most research (including this project) focuses on a commonly used subset of **14 attributes** (13 features + 1 target).

**Dataset summary**:

* **File Name**: heart.csv
* **Source**: Kaggle - Heart Disease UCI
* **Total Rows**: 1025
* **Total Columns**: 14
* **Missing Values**: None (All columns are clean)
* **Target Distribution**:
  + 1 → 526 patients with heart disease
  + 0 → 499 patients without heart disease

**FEATURES (ATTRIBUTES USED)**:

The dataset used for this heart disease prediction model contains 13 features that describe various medical conditions and personal attributes of patients. Below is a detailed explanation of each:

**Age**:

This represents the age of the patient in years. Heart disease risk tends to increase with age, making this an essential factor in prediction.

**Sex**:

This is a binary variable indicating the gender of the patient — 1 for male and 0 for female. Studies show that gender can influence the likelihood of developing heart conditions, with men generally at higher risk at earlier ages.

**Chest pain type (cp)**:

Chest pain is a key indicator of heart disease. This categorical feature has four possible values:

* 0: Typical angina (pain related to reduced blood flow to the heart)
* 1: Atypical angina (chest discomfort that doesn’t follow the usual pattern of angina)
* 2: Non-anginal pain (not related to heart)
* 3: Asymptomatic (no pain but disease might be present.

**Resting Blood Pressure (trestbps)**:

This numerical feature indicates the blood pressure in mm Hg when the patient is at rest. Elevated resting blood pressure is a common symptom of cardiovascular conditions.

**Serum Cholesterol (chol)**:

Measured in mg/dL, this represents the cholesterol level in the blood. High cholesterol can lead to blockages in arteries, increasing the risk of heart disease.

**Fasting Blood sugar (fbs)**:

This binary feature shows whether the fasting blood sugar level is greater than 120 mg/dL. A value of 1 indicates high blood sugar (possible diabetes), which is a risk factor for heart problems.

**Resting Electrocardiographic Results (restecg)**:

This categorical feature describes the results of the patient’s resting ECG:

* 0: Normal
* 1: ST-T wave abnormality (may indicate heart issues)
* 2: Left ventricular hypertrophy (thickening of the heart wall)

**Maximum Heart Rate Achieved (thalach)**:

This numerical value measures the highest heart rate achieved during exercise. A lower than expected maximum heart rate can signal underlying heart issues.

**Exercise-Induced Angina (exang)**:

This binary feature indicates whether the patient experienced angina (chest pain) during physical activity — 1 for yes, 0 for no. The presence of exercise-induced angina often suggests coronary artery disease.

**ST Depression (oldpeak)**:

This feature represents the amount of depression in the ST segment of the ECG after exercise, compared to rest. A higher value suggests greater heart stress and higher risk.

**Slope of the Peak Exercise ST Segment(slope)**:

This categorical feature indicates the shape of the ST segment during peak exercise:

* 0: Upsloping
* 1: Flat
* 2: Downsloping  
   A flat or downsloping ST segment often indicates abnormalities.

**Number of Major Vessels (ca)**:

This indicates how many major vessels (ranging from 0 to 3) were visualized using fluoroscopy. Higher numbers can signify more significant blockages or abnormalities in blood flow.

**Thalassemia (thal)**:

This describes the result of a thallium stress test related to blood disorders. The values are:

* 1: Fixed defect (permanent damage)
* 2: Normal
* 3: Reversible defect (temporary stress-related abnormality)  
   These outcomes help assess the heart’s blood flow under stress conditions.

**Target**:

* 0: No heart disease (the patient does not exhibit clinical signs of heart disease based on the diagnostic features provided)
* 1: Heart disease present (the patient is likely to have some form of heart disease as indicated by patterns in the medical attributes)

**DATA PROCESSING**

Before training the machine learning model, the dataset underwent several preprocessing steps to ensure data quality and optimal model performance. These steps are crucial to handle inconsistencies, convert data into a usable format, and enhance prediction accuracy.

**Loading the Dataset**:

The dataset was loaded using the pandas library in Python from a CSV file named heart.csv. The structure and contents were first examined using functions like .head(), .info(), and .describe().

**Handling missing values**:

The dataset was checked for missing values using df.isnull().sum(). Fortunately, this dataset contained **no missing values**, so no imputation or removal was necessary.

**Encoding categorical values**:

Some categorical features such as sex, cp, fbs, restecg, exang, slope, ca, and thal were already **numerically encoded**, so no further encoding was needed. These were treated as categorical during model training.

**Feature selection**:

Only the **13 most relevant features** were used from the dataset to train the model. The target column was designated as the **dependent variable**, while the remaining 13 columns were considered **independent features**.

**Train-Test Split**:

The dataset was split into **training and testing sets** using scikit-learn’s train\_test\_split() function. Typically, 70–80% of the data was used for training and the rest for testing, ensuring that the model could generalize well to unseen data.

**Model serialization**:

After training the model, it was saved using the pickle module. This allows the trained model (model\_heart.pkl) to be loaded directly in the Streamlit app without retraining.

**Web Application Interface**

The web application was built using **Streamlit**, a lightweight and intuitive Python framework for creating interactive web apps with minimal code.

**Key Features**:

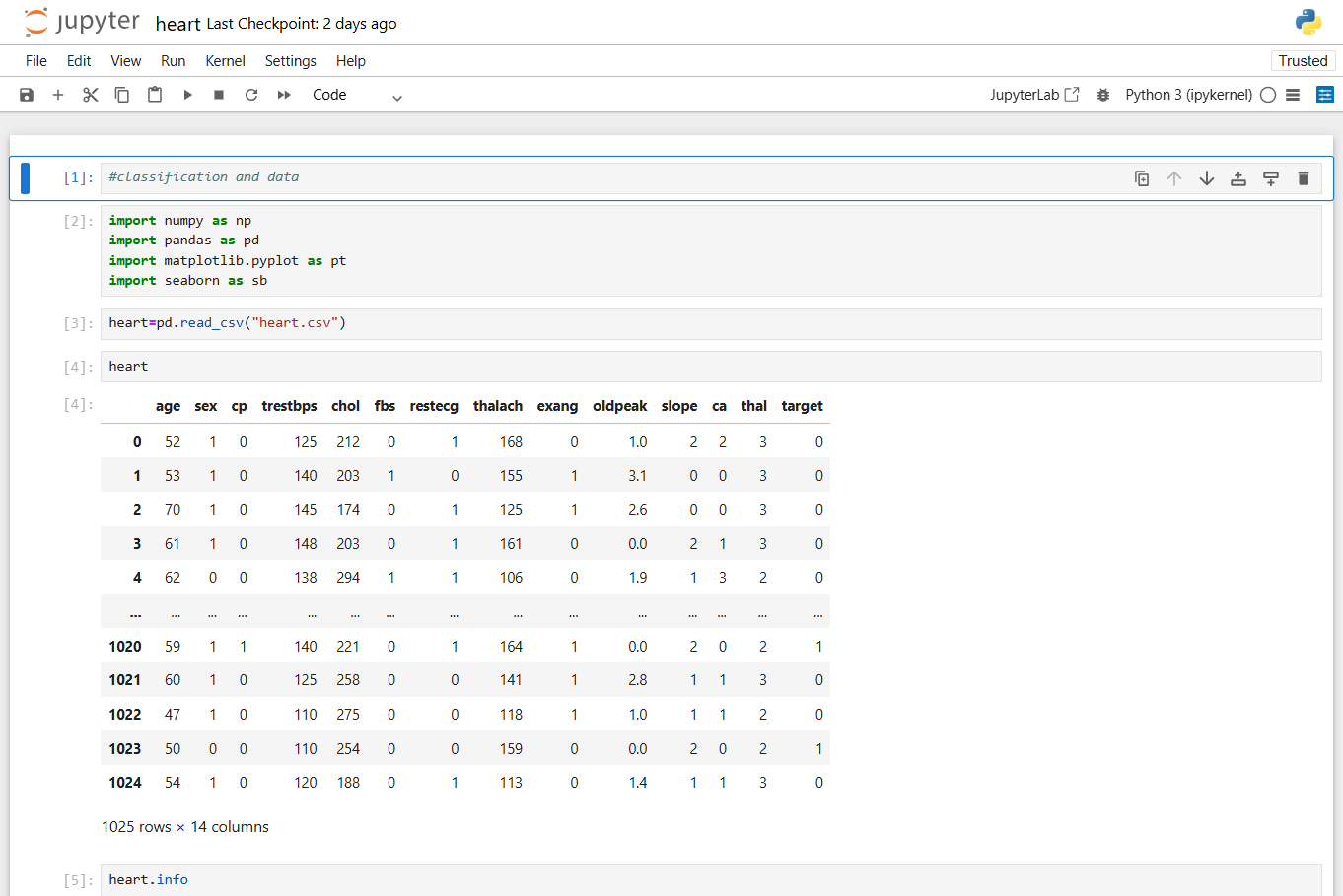
* **User Input Form**: Users can enter relevant medical information such as age, gender, cholesterol level, chest pain type, and more using text boxes and radio buttons.
* **Prediction Button**: A simple **"Detect the Heart Disease"** button runs the prediction logic using the trained model.
* **Result Display**:
  + If heart disease is predicted: Displays an error message with ❄️ st.snow() effect.
  + If no heart disease is detected: Shows a success message with 🎈 st.balloons() effect.

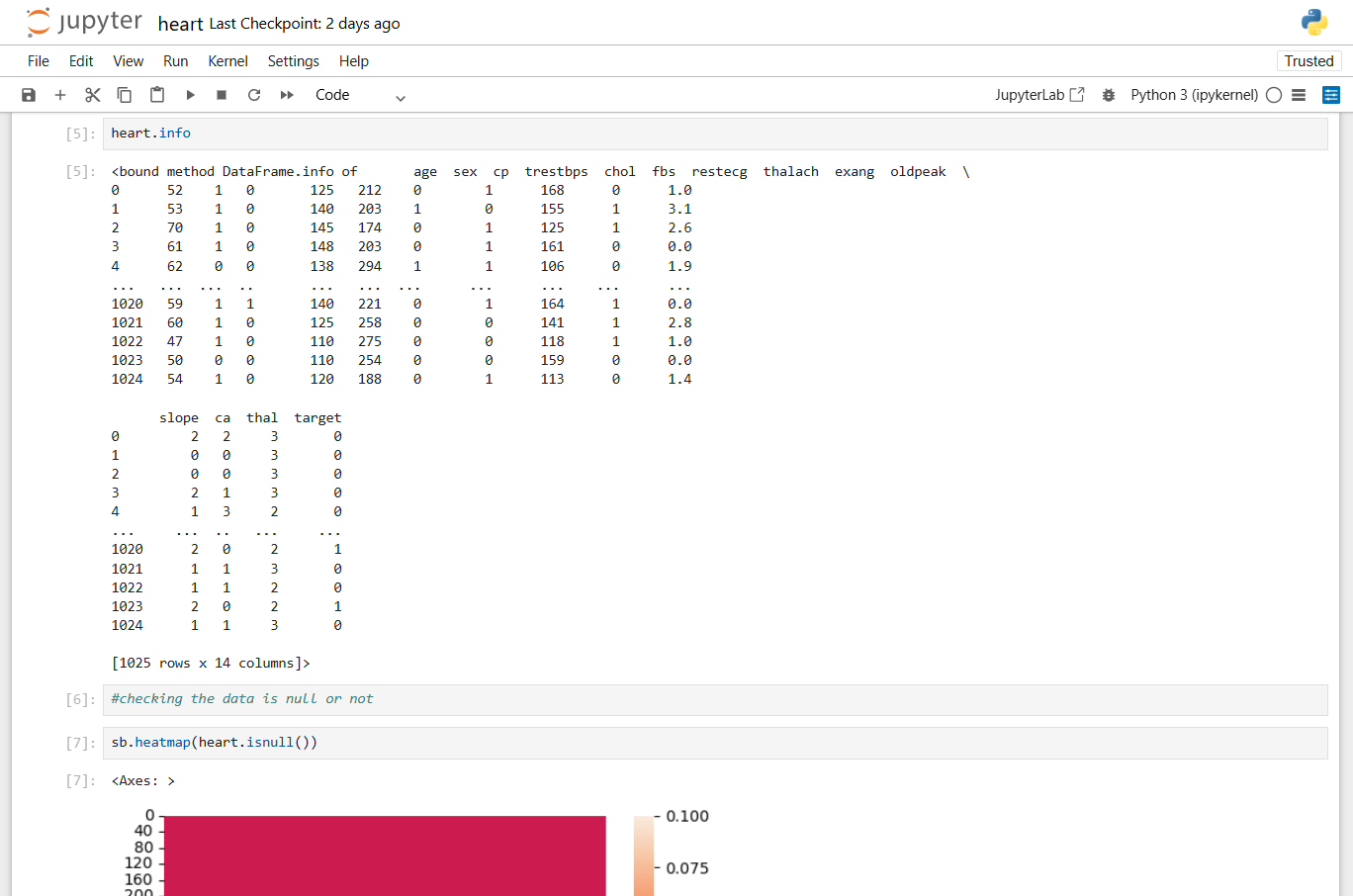
**Streamlit Widgets used**:

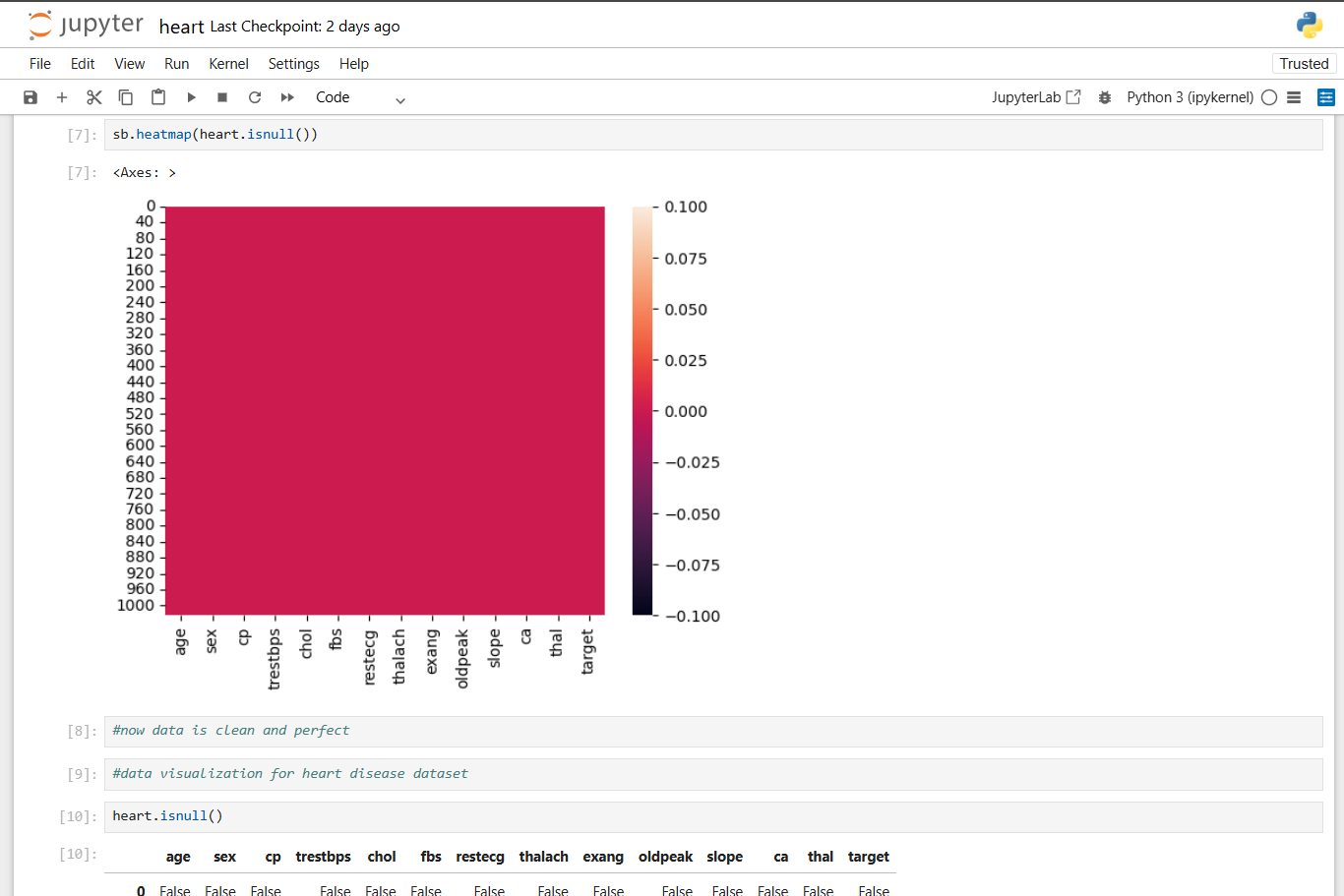
* st.text\_input() – for numeric user inputs (e.g., age, cholesterol)
* st.radio() – for categorical selections (e.g., sex, chest pain type, thalassemia)
* st.button() – to trigger model prediction
* st.success() and st.error() – to visually display the result
* st.balloons() and st.snow() – for engaging feedback based on prediction

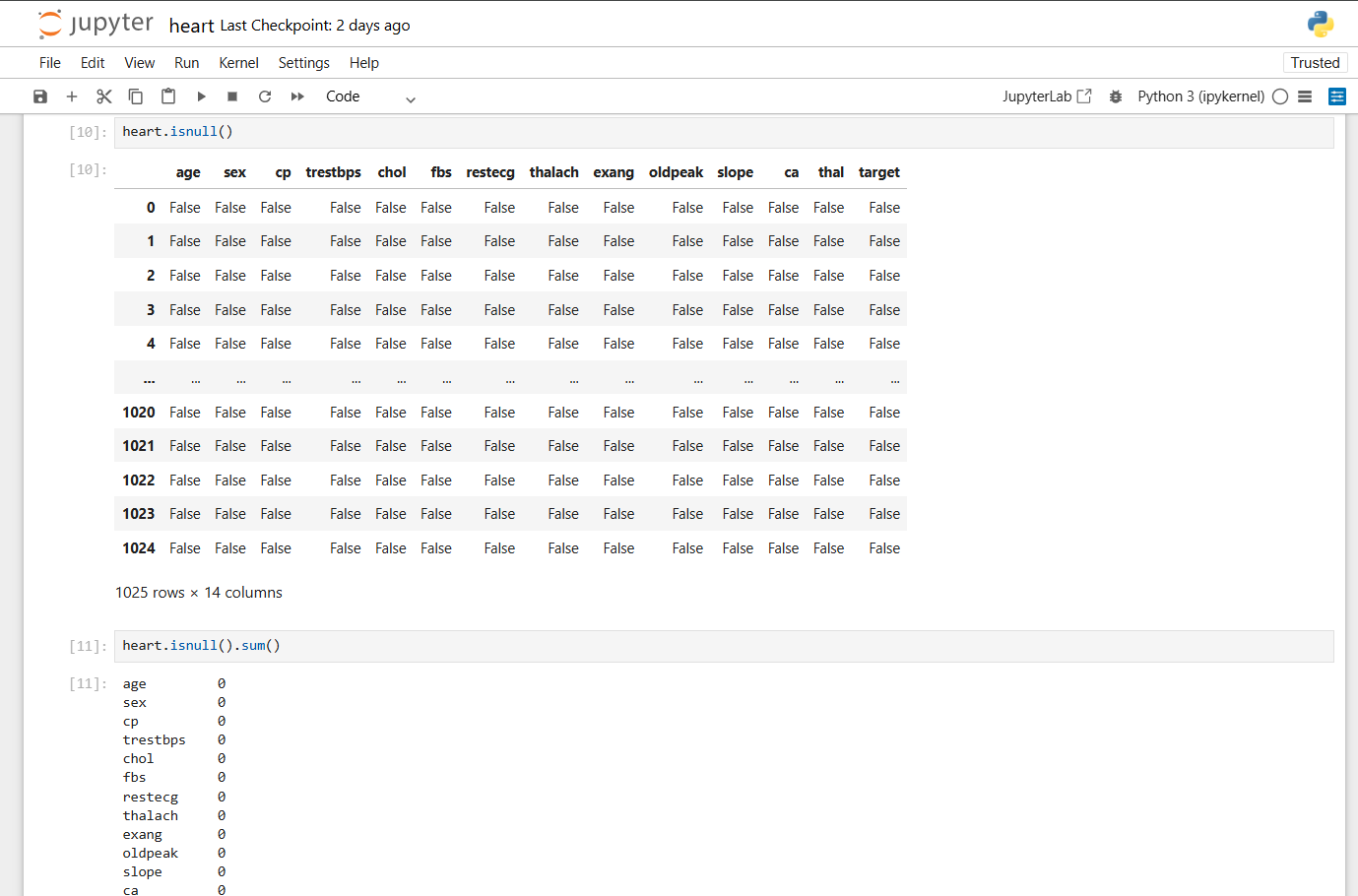
**RESULT AND OUTPUT**

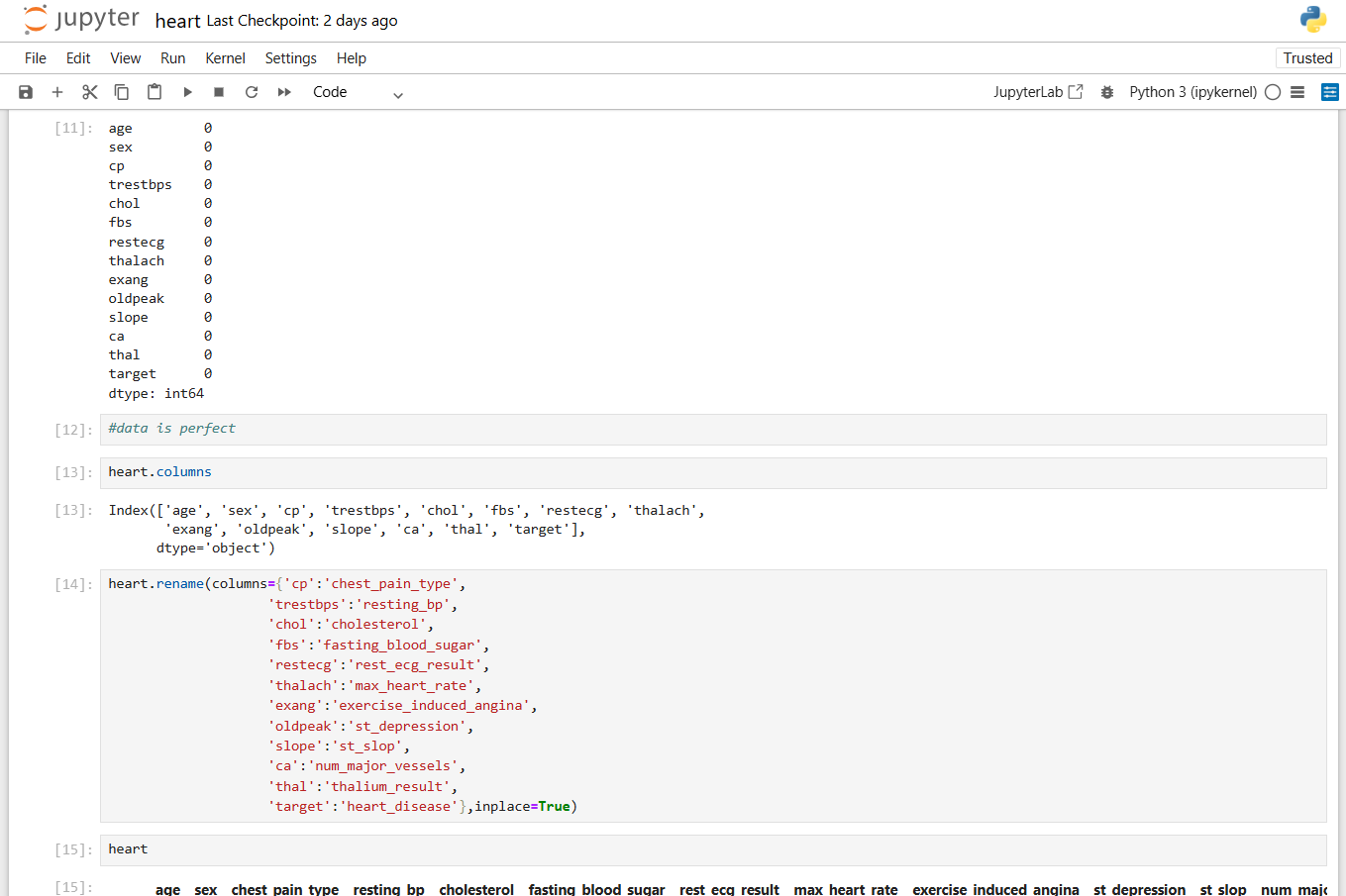
**Code:**

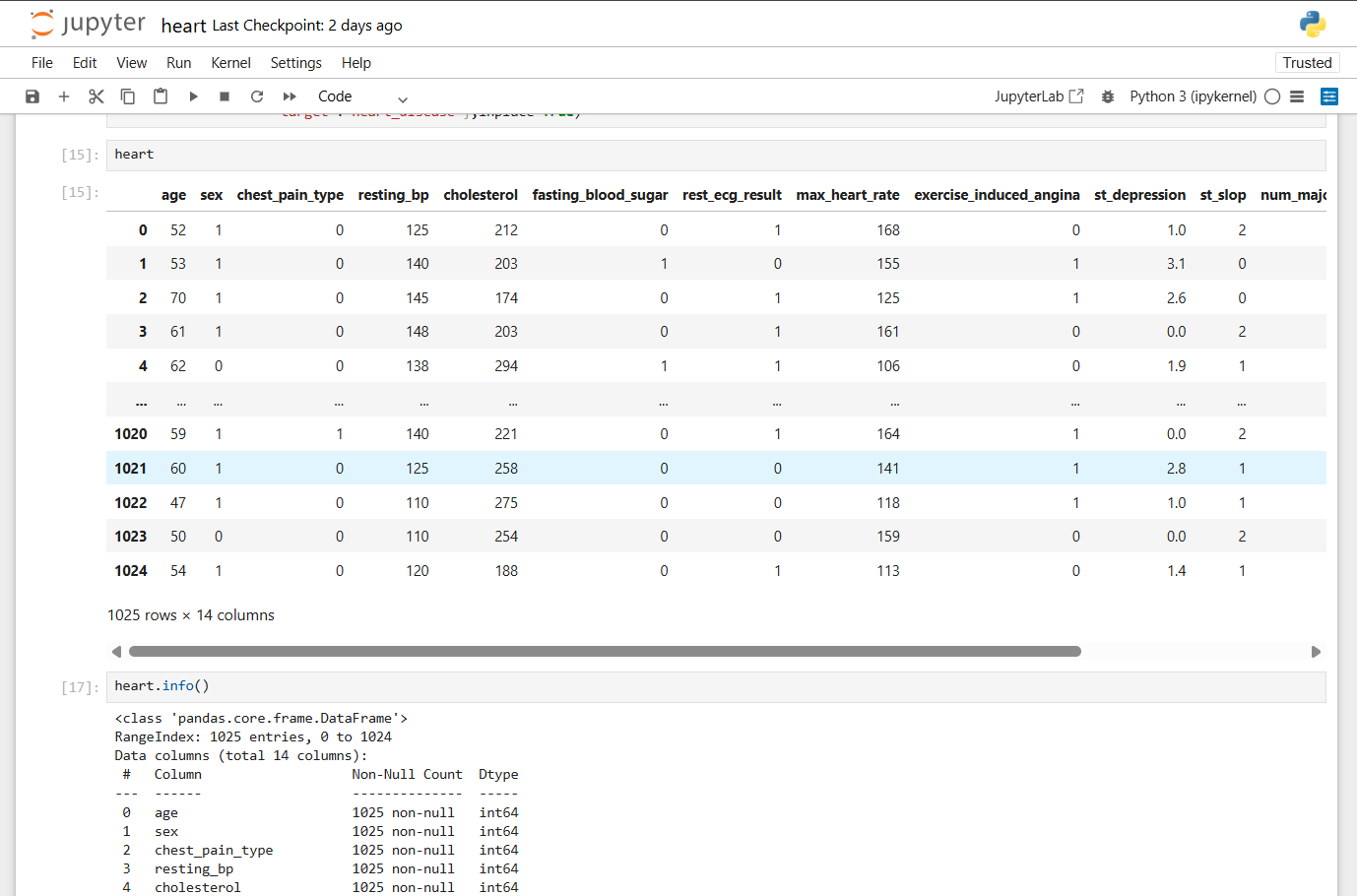


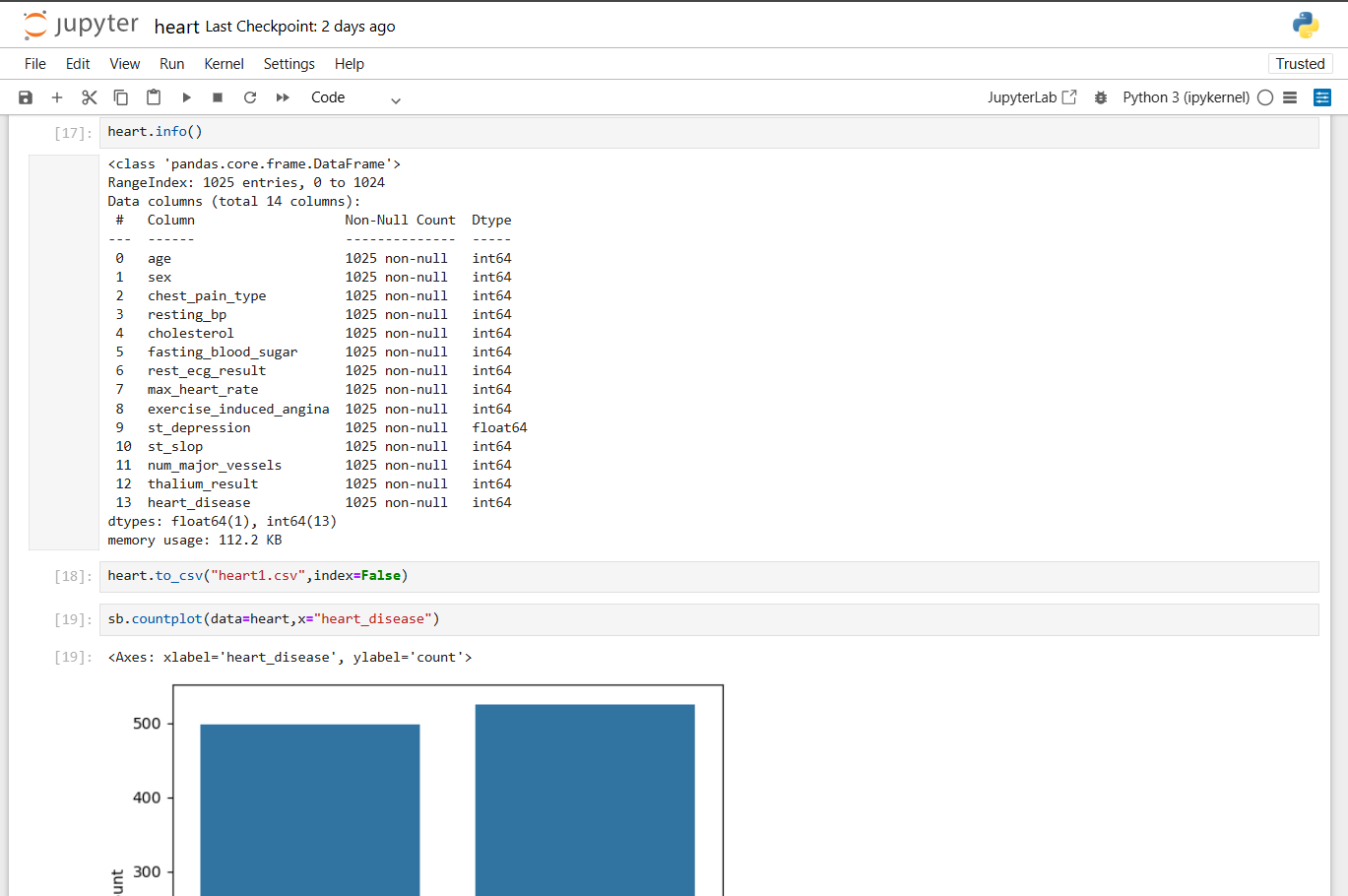


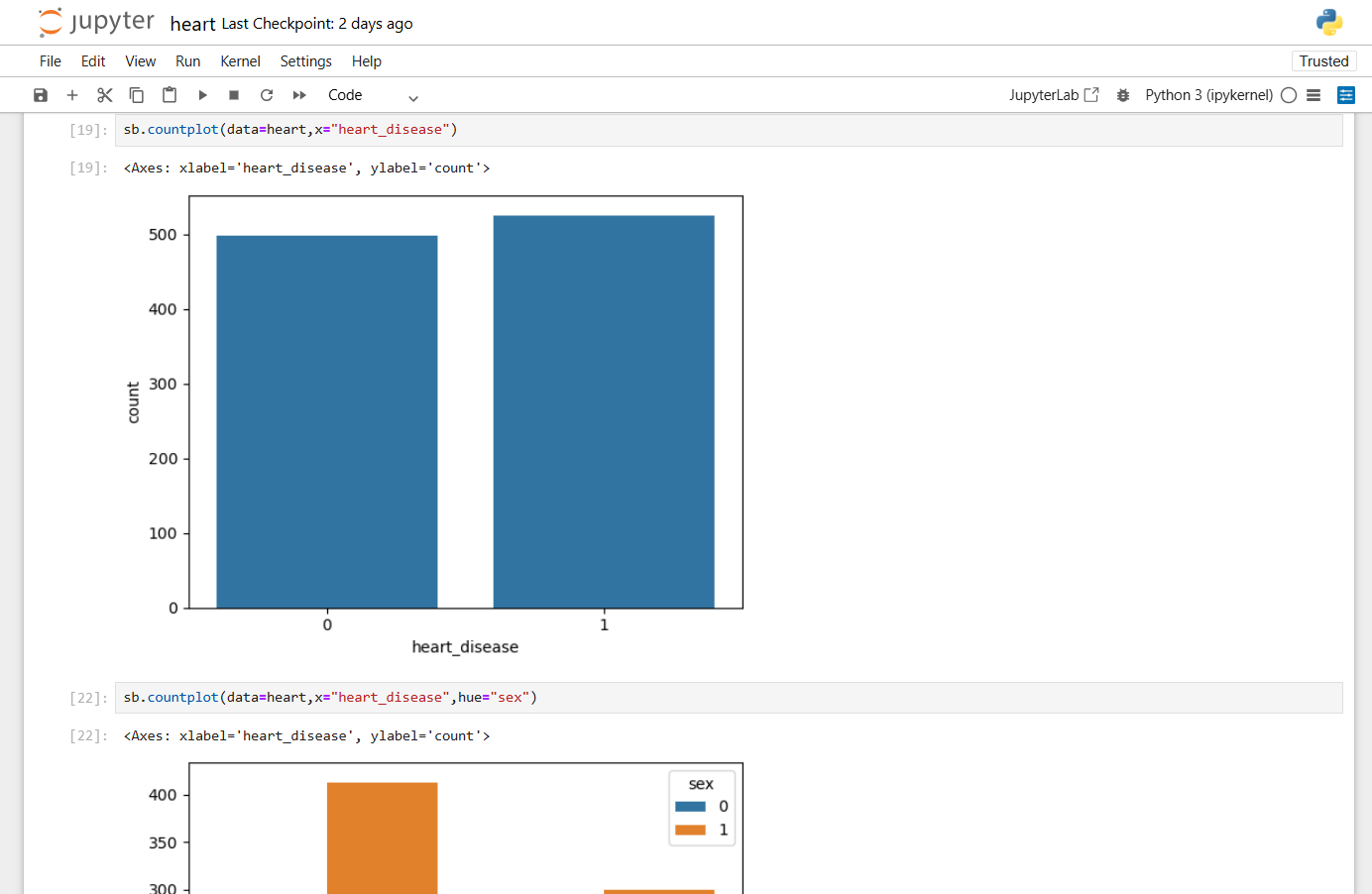


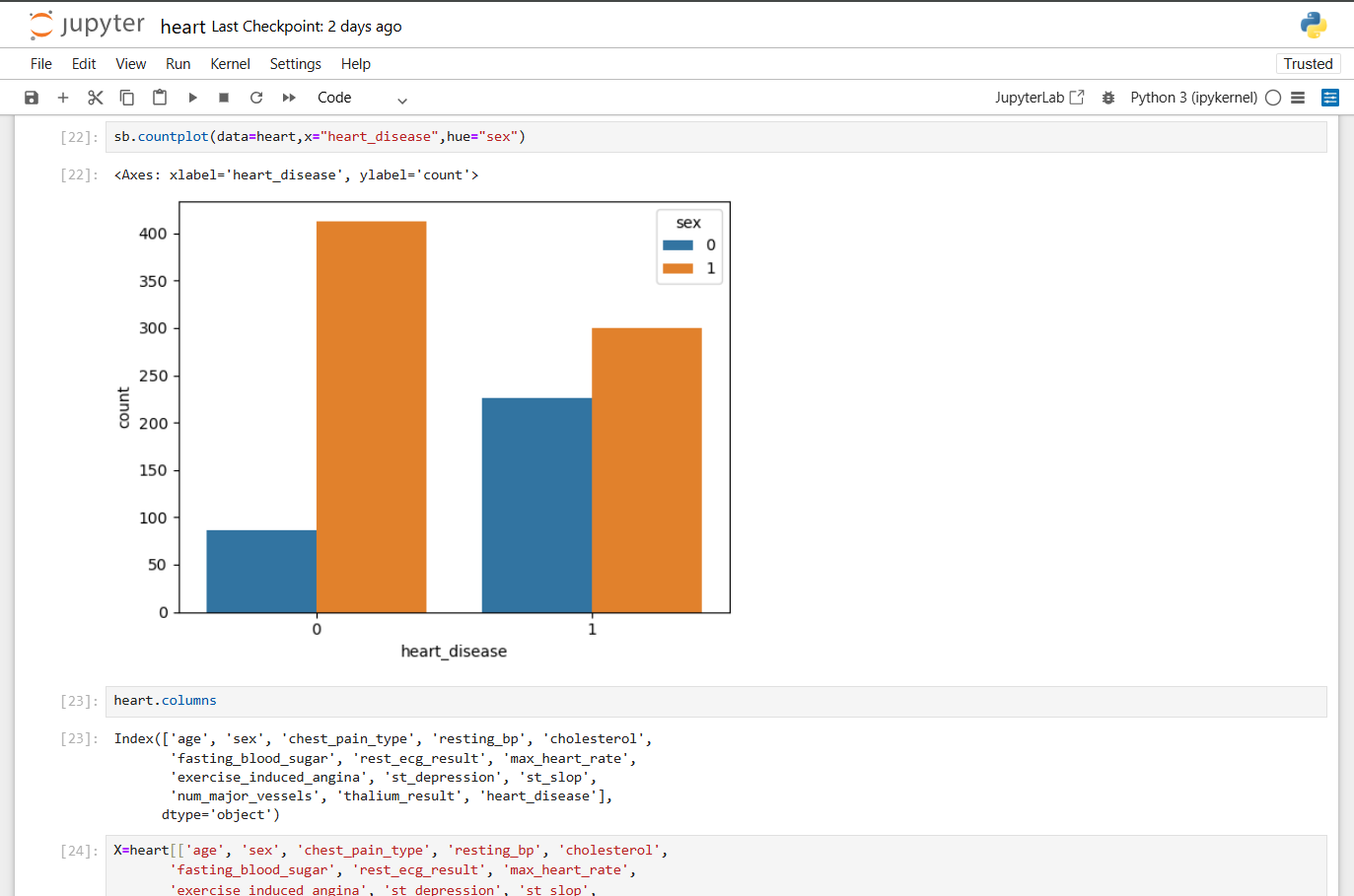


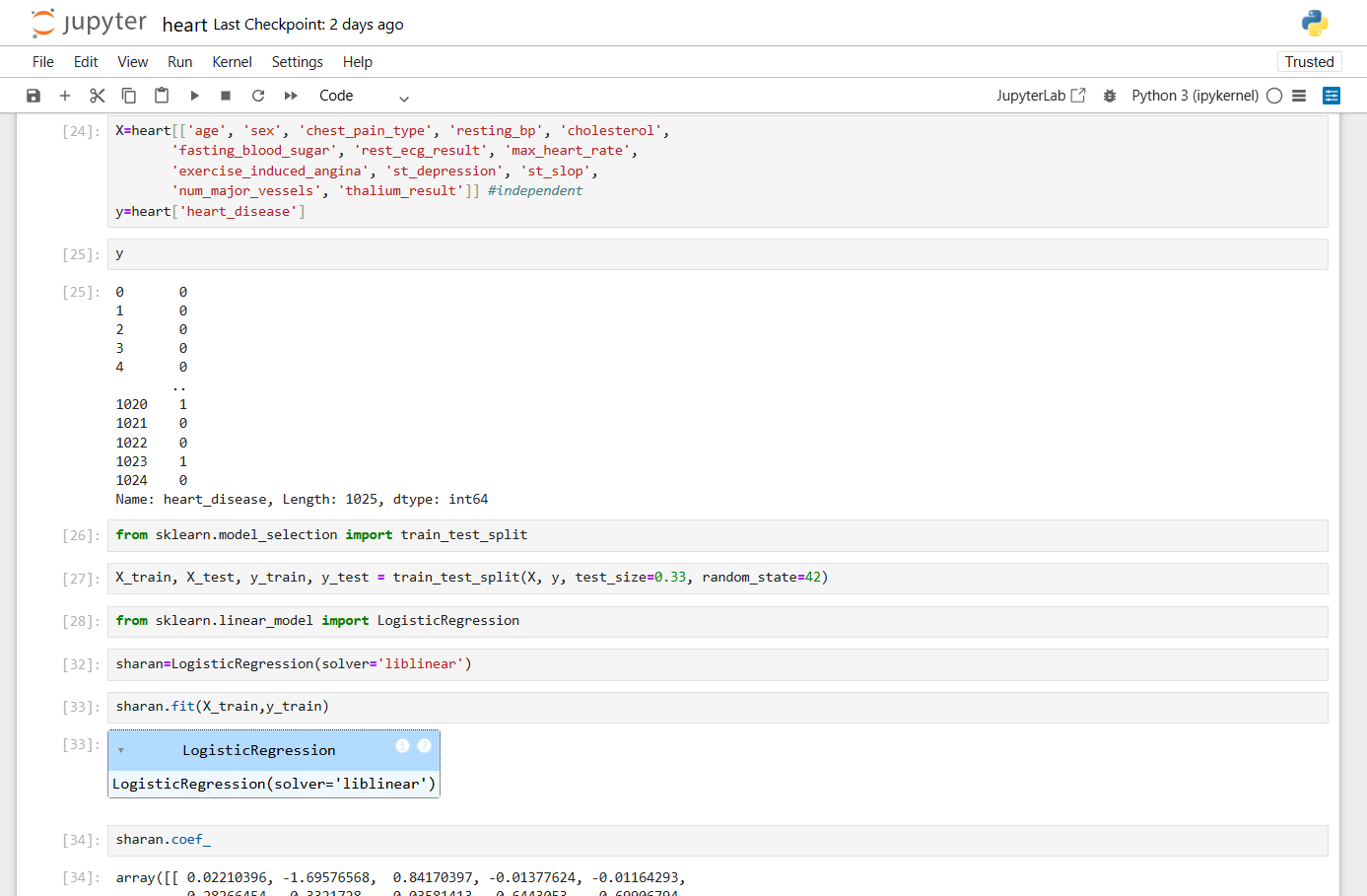


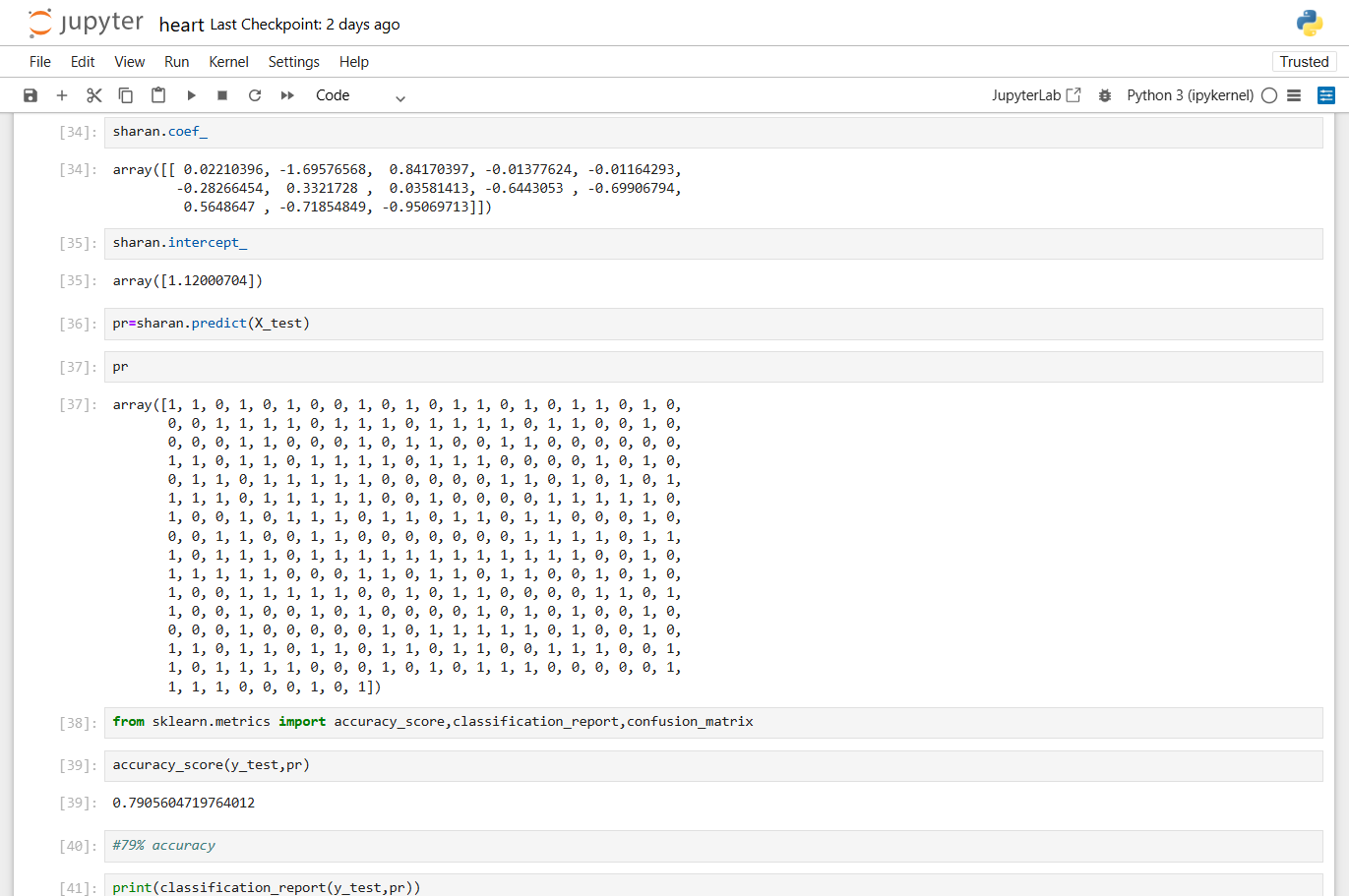


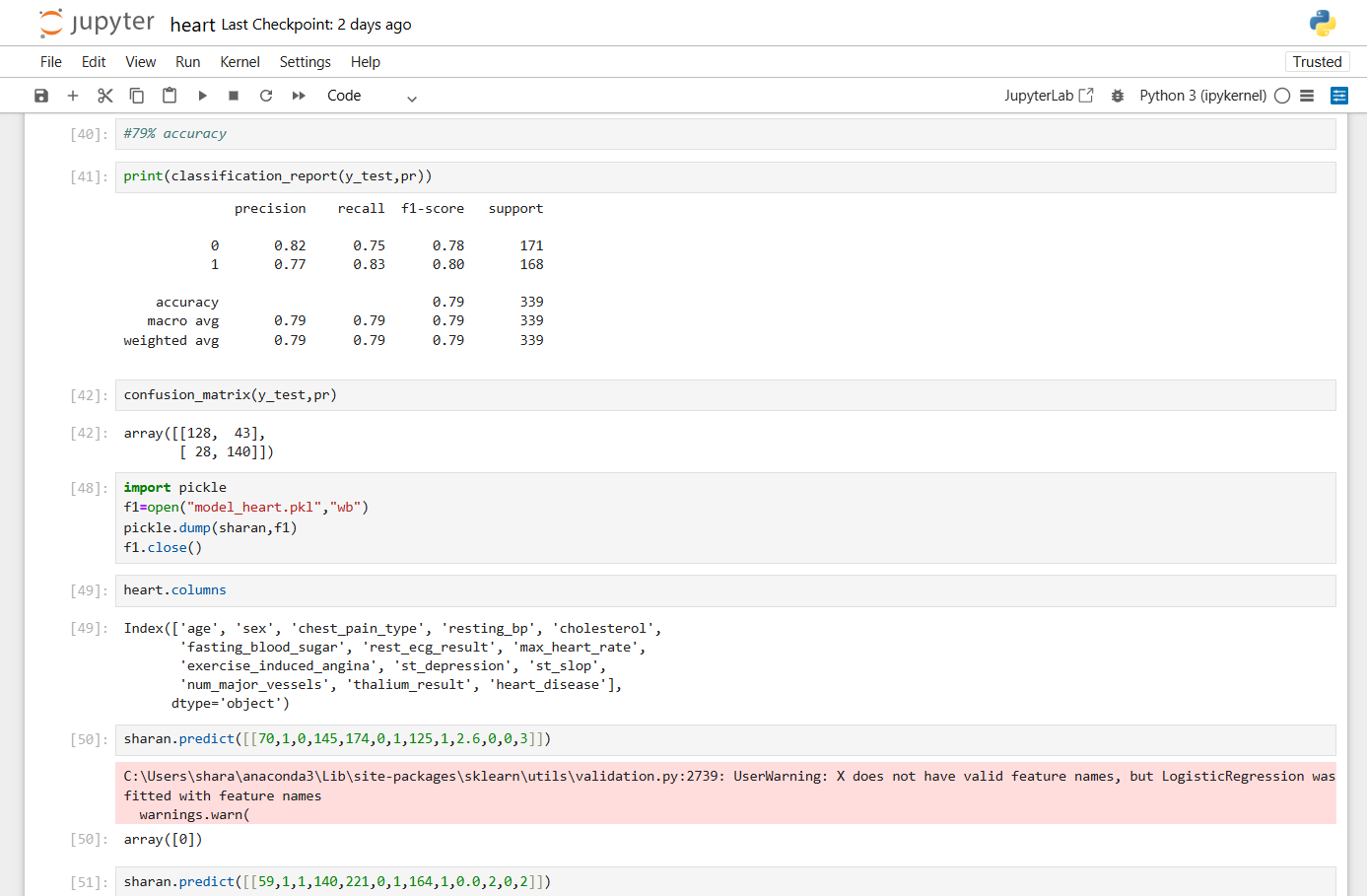


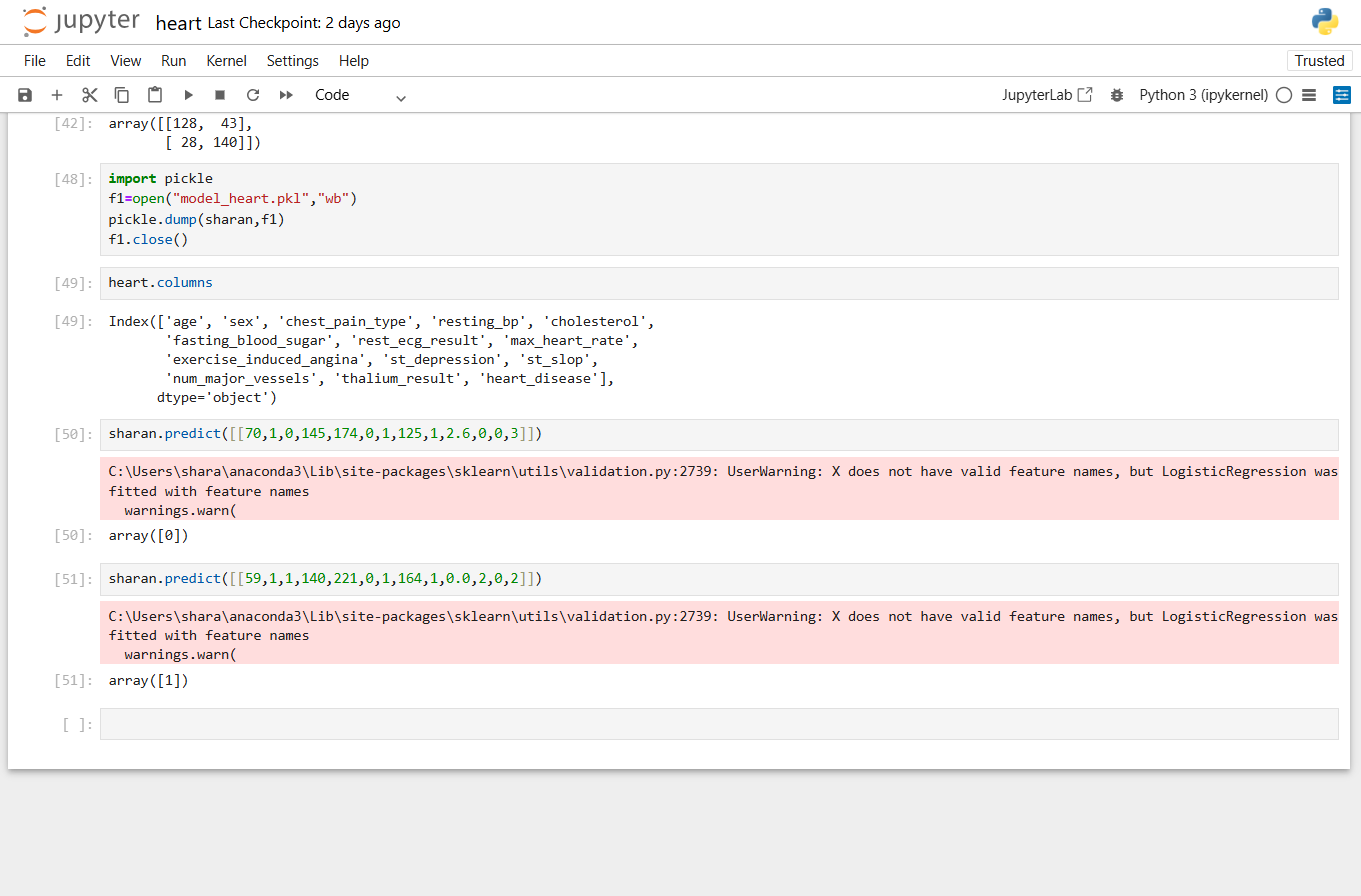




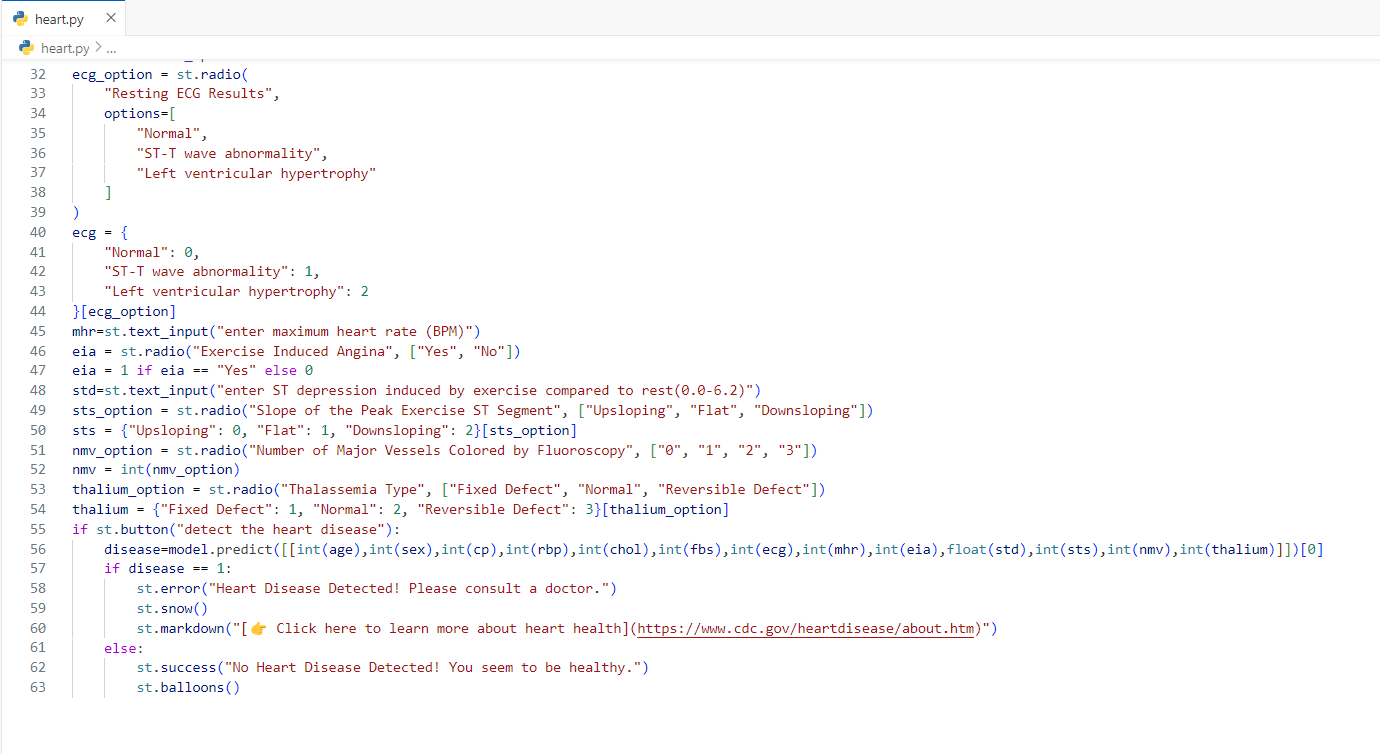


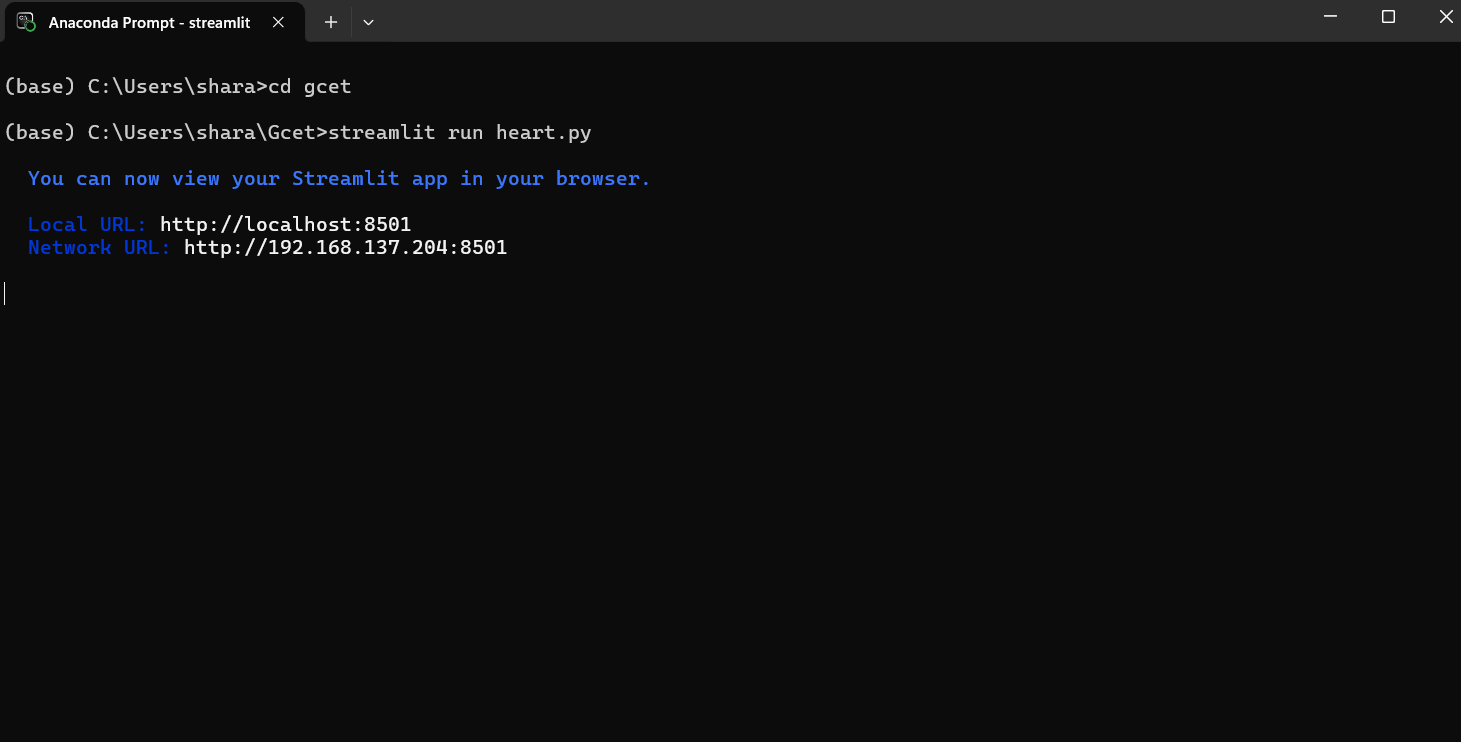




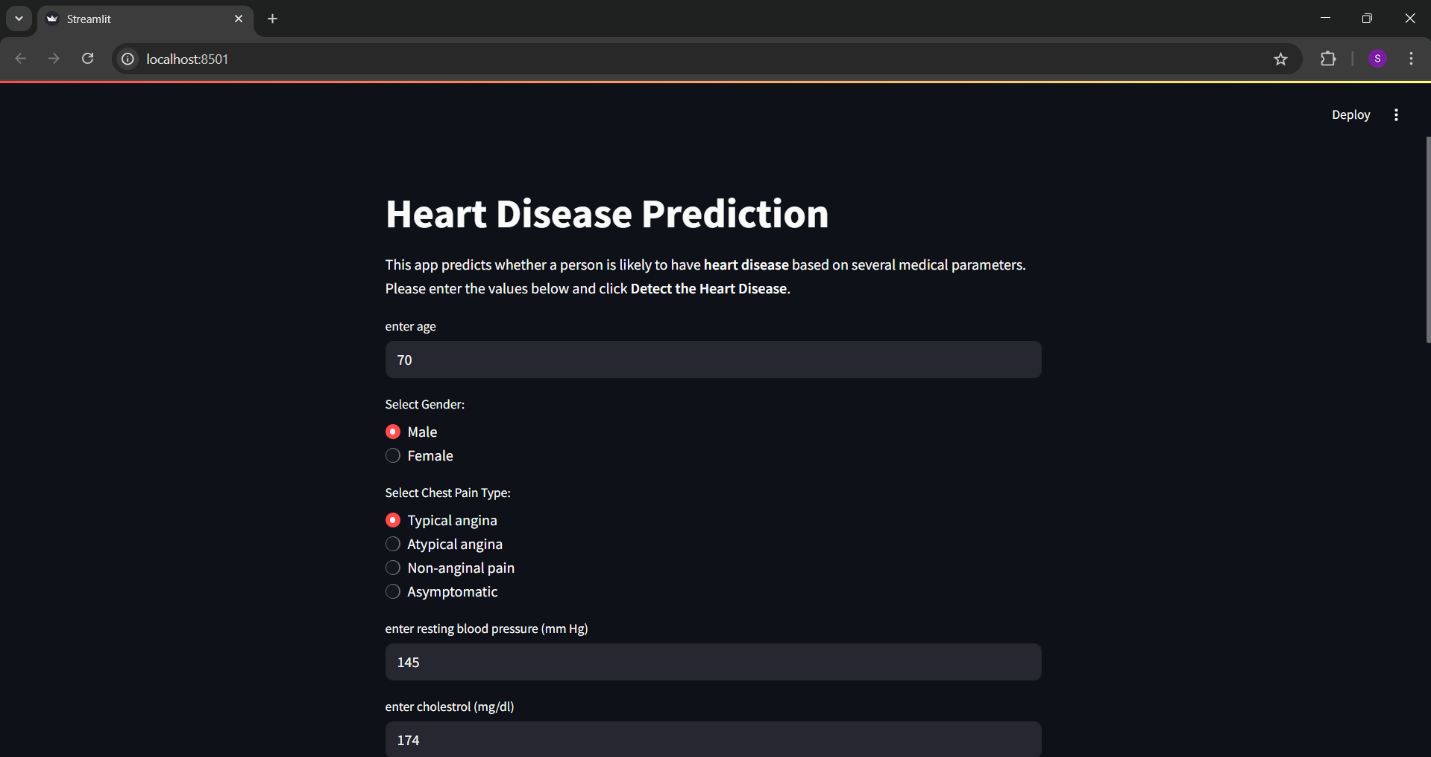


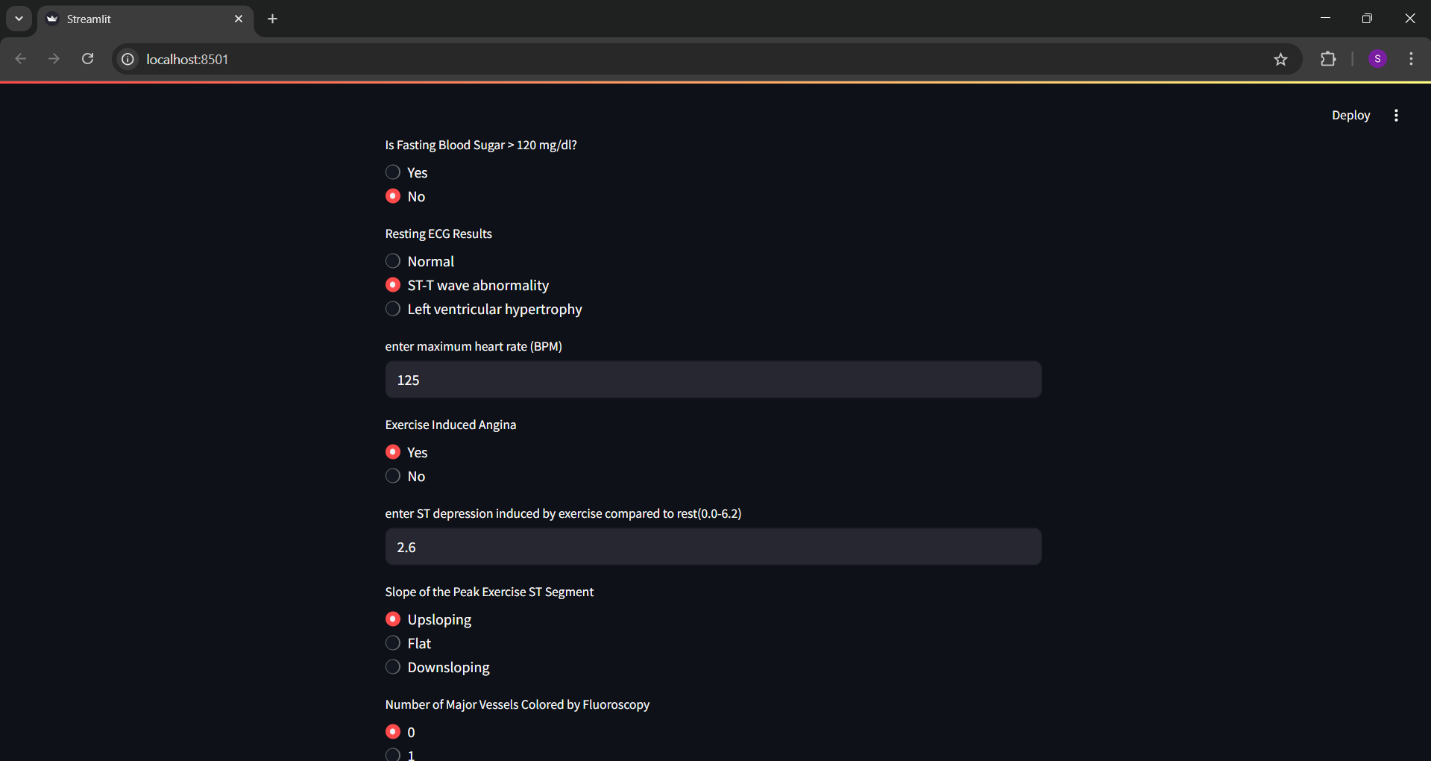


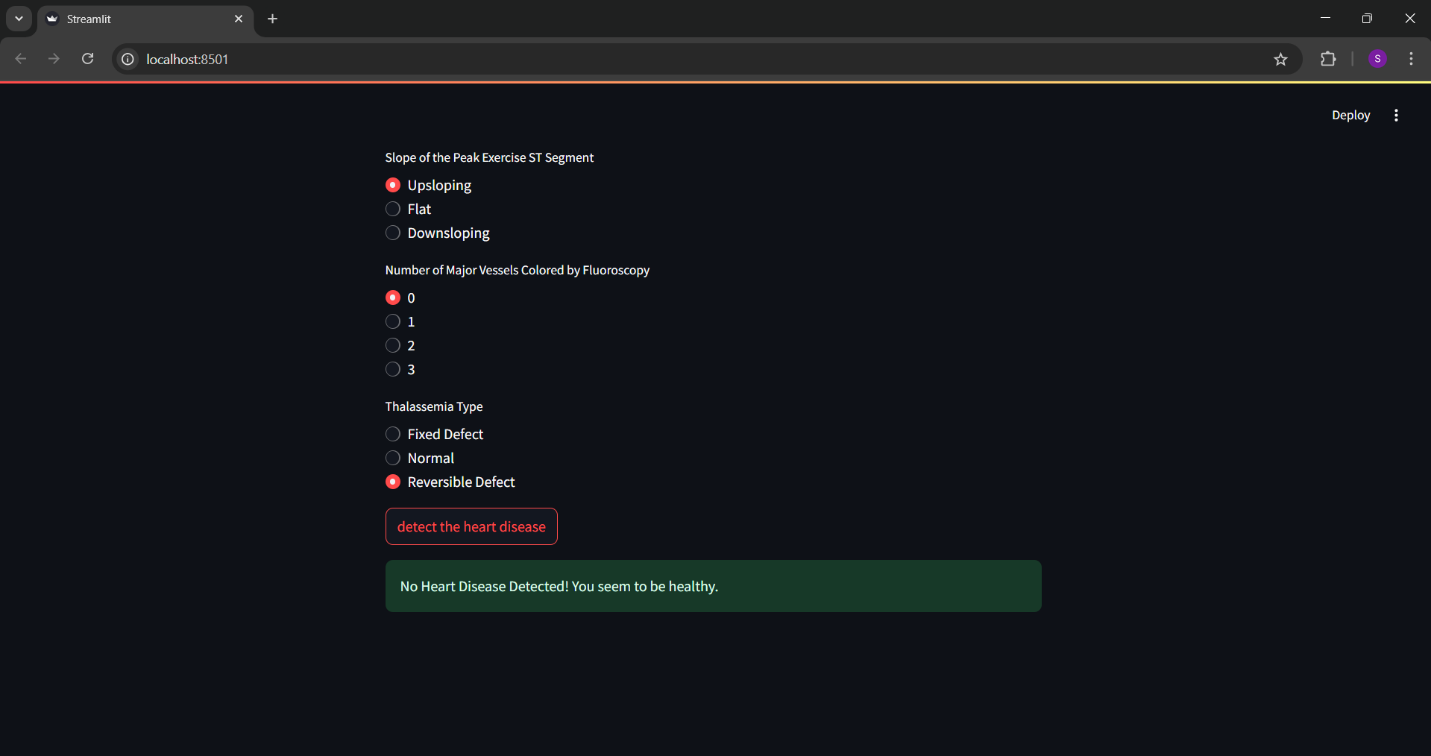




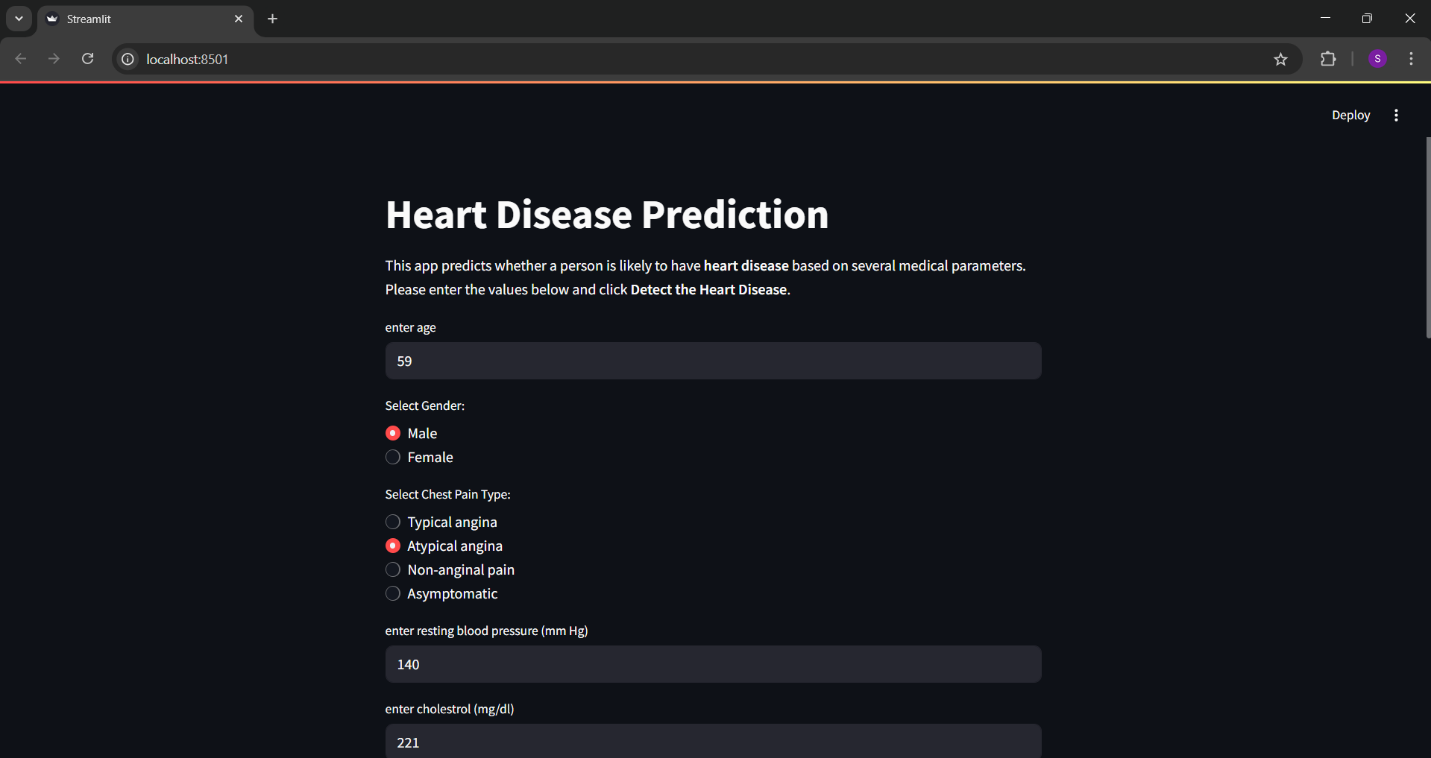
**OUTPUT FOR A PERSON WITH NO HEART DISEASE**:

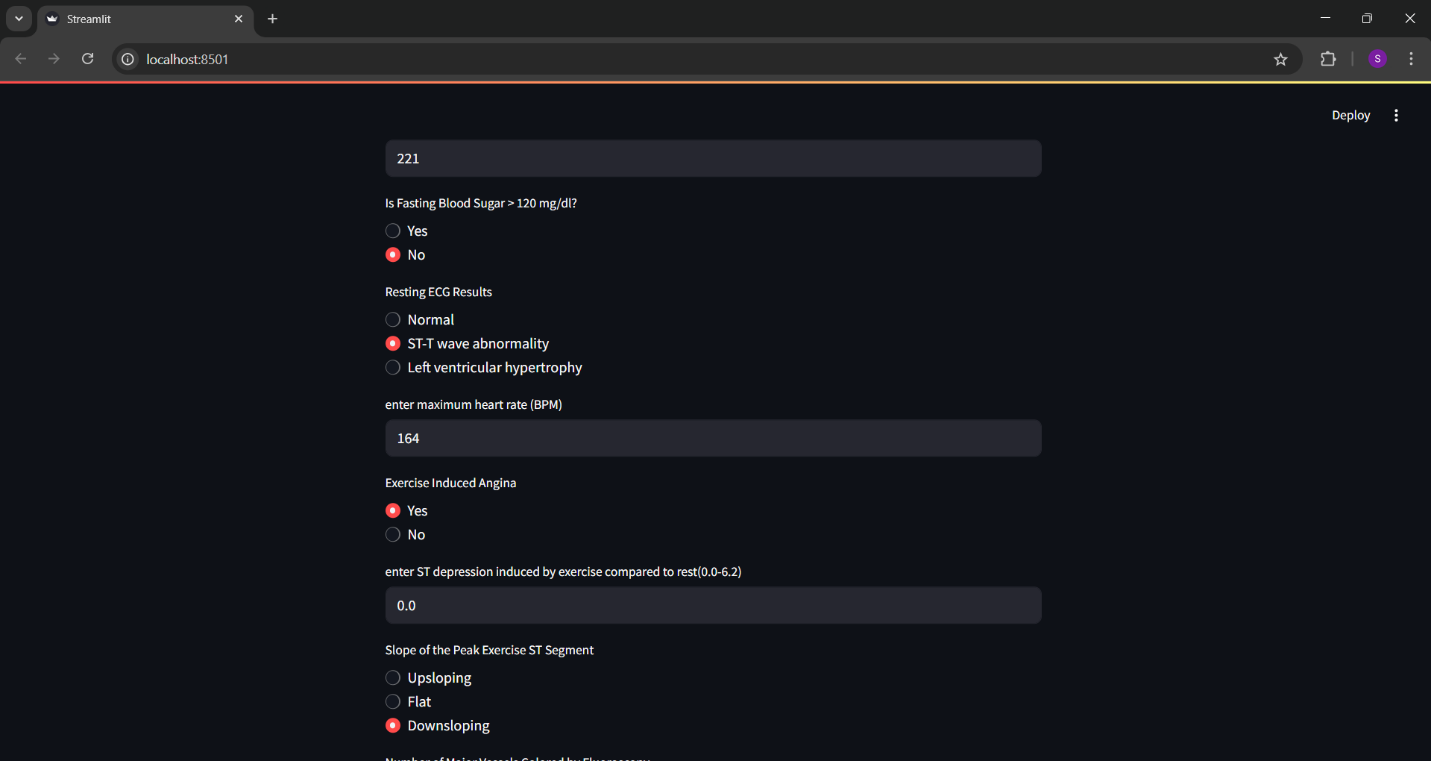


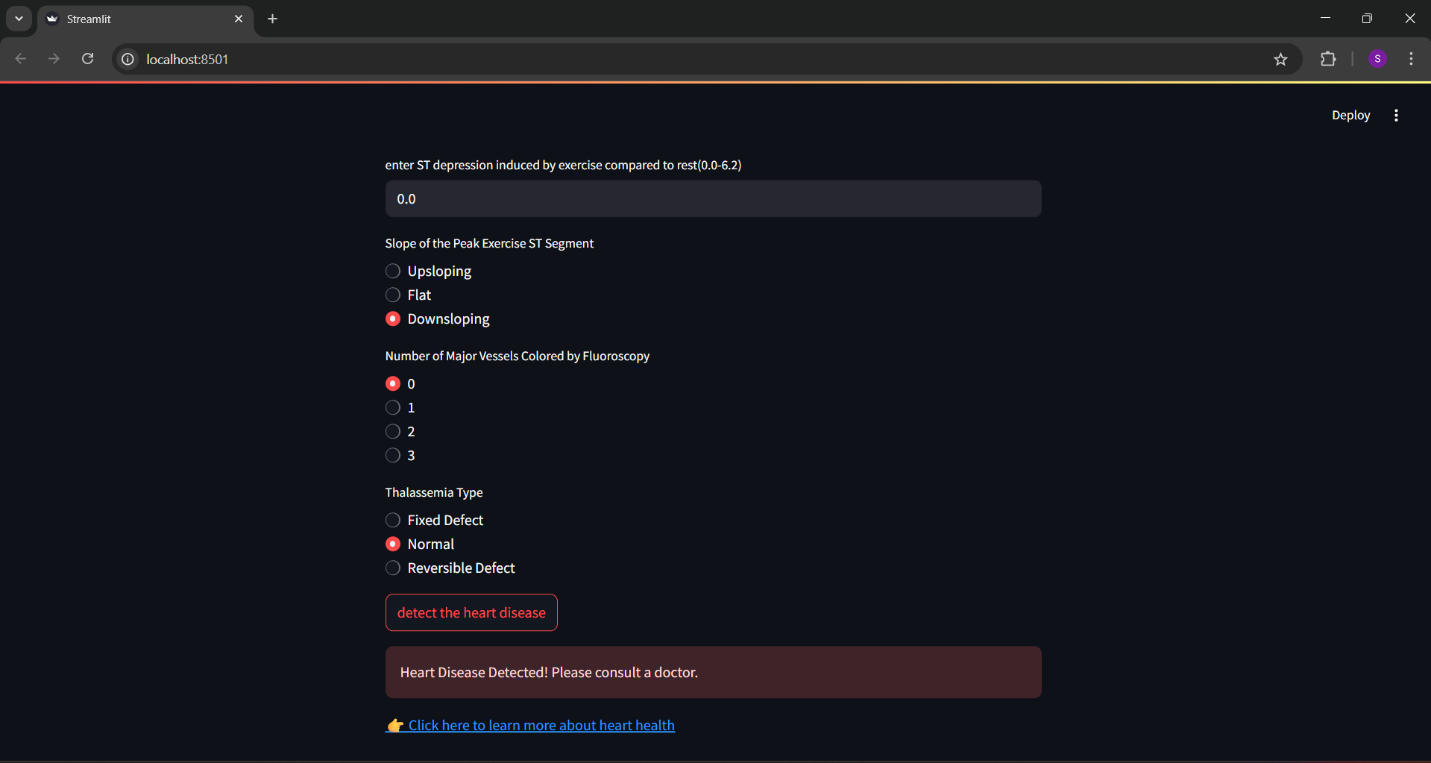


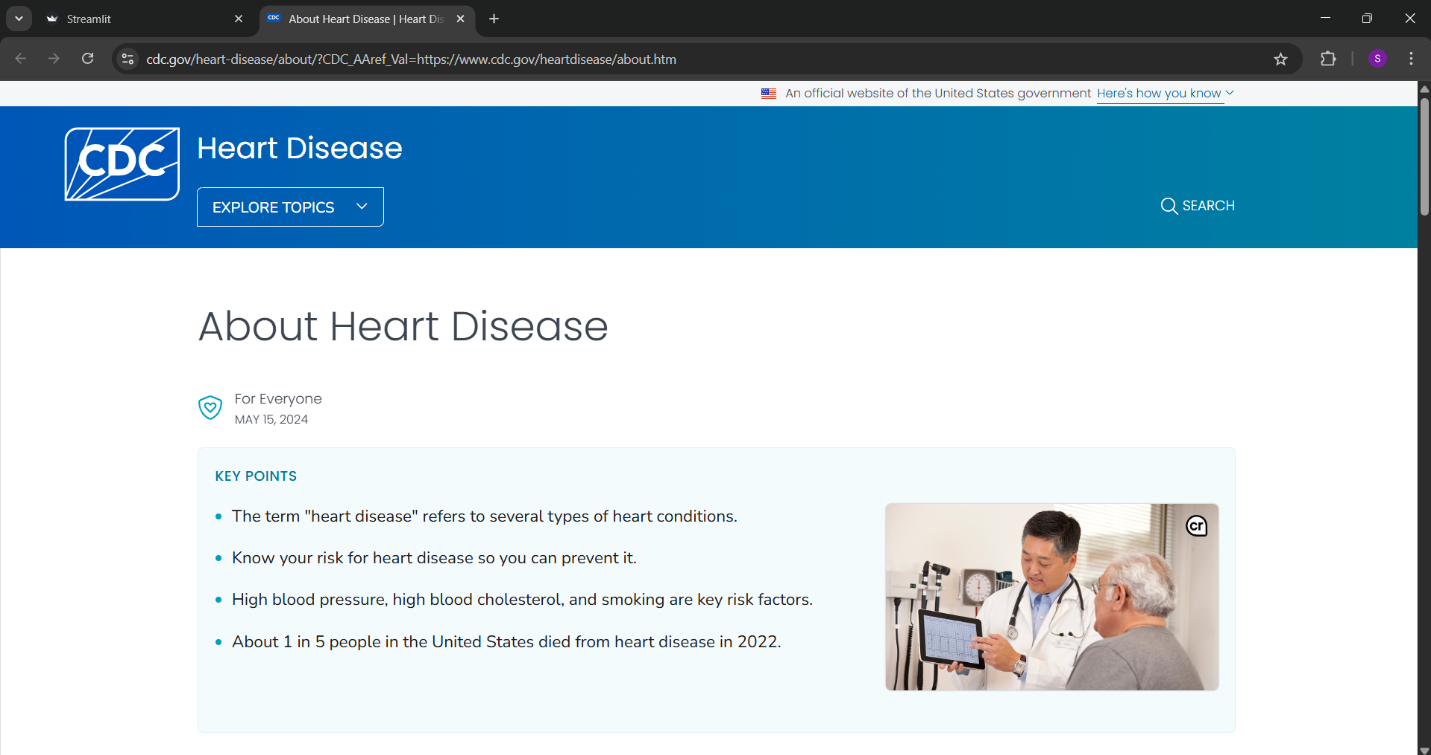


**OUTPUT FOR A PERSON WITH HEART DISEASE**:









**CONCLUSION**

The heart disease prediction project demonstrates how machine learning can be effectively applied in healthcare for **early diagnosis and awareness**. By using a trained ML model and deploying it via a web application, the system allows users to assess their risk of heart disease based on simple input parameters.

The application is fast, lightweight, and easy to use, making it a useful tool for basic health screening. Although the tool is not a substitute for clinical diagnosis, it serves as a helpful first step toward better heart health awareness.

This project successfully combines **data science**, **machine learning**, and **interactive UI design** to deliver a functional and meaningful solution to a real-world problem.

**REFERENCES**

* **Kaggle - Heart Disease Dataset by johnsmith88**  
   One of the most widely used datasets for heart disease prediction projects, containing labeled clinical data of patients.  
   Link: <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>
* **scikit-learn Documentation**  
   Used for building and evaluating the machine learning model.  
   Link: [https://scikit-learn.org](https://scikit-learn.org/)
* **Streamlit Documentation**  
   Used for developing the interactive web application.  
   Link: <https://docs.streamlit.io>
* **Python Official Documentation**  
   Reference for Python syntax and standard libraries.  
   Link: <https://docs.python.org/3/>
* **CDC – Heart Disease Information**  
   Helped understand the importance and medical context of heart disease.  
   Link: <https://www.cdc.gov/heartdisease>