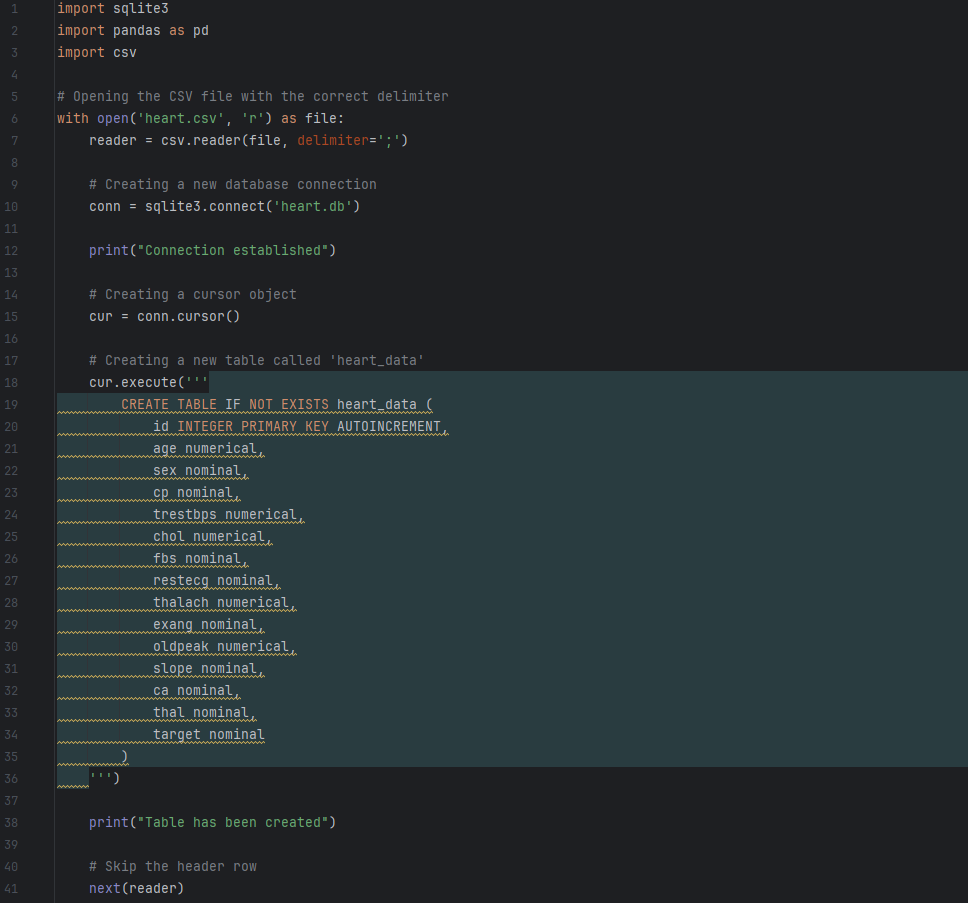
**Section A**

**Question 1**

Connecting to a SQLite database.

(What is SQLite (DB Create and Connection) & How to Create Table in Python, 2022). (How To Import Csv Datasets in Python Pandas, 2019) (how to read from csv file in python |csv reader, csv DictReader python|csv.reader vs csv.dictreader, 2021)

A screenshot of a computer

Description automatically generatedA screen shot of a computer

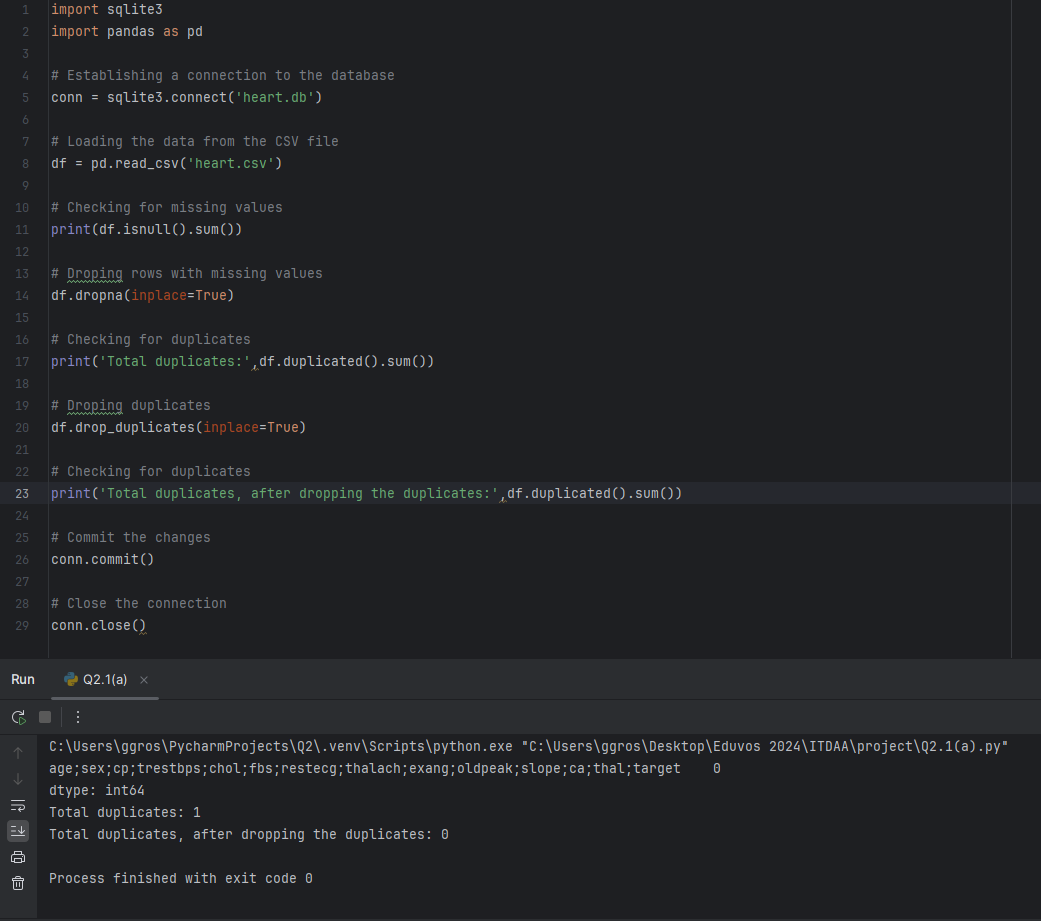
Description automatically generated (How to connect Python with a SQL database [SQLITE] | perfect for beginners, 2021) (SQLite Databases With Python - Full Course, 2020)A screenshot of a computer

Description automatically generated (What is SQLite (DB Create and Connection) & How to Create Table in Python, 2022)

**Question 2**

2.1 Preprocessing and visualizing the data

a. Perform any necessary cleaning and preprocessing of the data.

 (Data Cleaning Tutorial | Cleaning Data With Python and Pandas, 2020)

b. Plot the distribution of classes for the (8) categorical variables based on the target variable. Provide any observations that can be derived from these plots.

A screenshot of a computer

Description automatically generated (Plotting Categorical Variables with matplotlib, 2022) (psu.edu, 2024)

The following observations were made from these plots:

1. Sex Distribution

- The distribution of sex for class 0 (no heart disease) is skewed toward males, while the distribution for class 1 (heart disease) is more balanced between males and females. This suggests that males are more likely to not have heart disease, while females are more likely to have heart diseases.

2. Chest Pain Distribution

- The distribution of chest pain types for class 0 is dominated by type 1 (typical angina), while the distribution of class 1 is more spread out across different types of chest pain. This suggest that people with heart disease are more likely to experience different types of chest pain.

3. Fasting blood sugar distribution

- The distribution of fbs for class 0 is skewed towards 0 (less than 120mg/dl), while the distribution for class 1 is more balanced between 0 and 1 (greater than 120mg/dl). This suggests that people with heart disease are more likely to have higher fasting blood sugar levels.

c. Plot the distribution of classes for the numeric variables based on the target variable. Provide anyobservations (at least 5) that can be derived from these plots.

A screenshot of a computer

Description automatically generated (Talks, YouTube)

The following observations were made from these plots:

1. Age Distribution

- The distribution of age for class 0 (no heart disease) is slightly skewed to the left, indicating that most people without heart diseases are younger. In contrast, the distribution of age for class 1 (heart disease) is more spread out, suggesting that heart diseases can affect people of all ages.

2. Tresrbps Distribution

- The distribution of tresbps (resting blood pressure) for class 0 is centered around 120-130, while class 1 is centered around 130-140, suggesting that people with heart disease tend to have higher resting blood pressure.

3. Chol Distribution

- The distribution of chol (serum cholesterol) for class 0 is centered around 200-220, while class 1 is centered around 220-240. This suggest that people with heart diseases tend to have higher serum cholesterol levels.

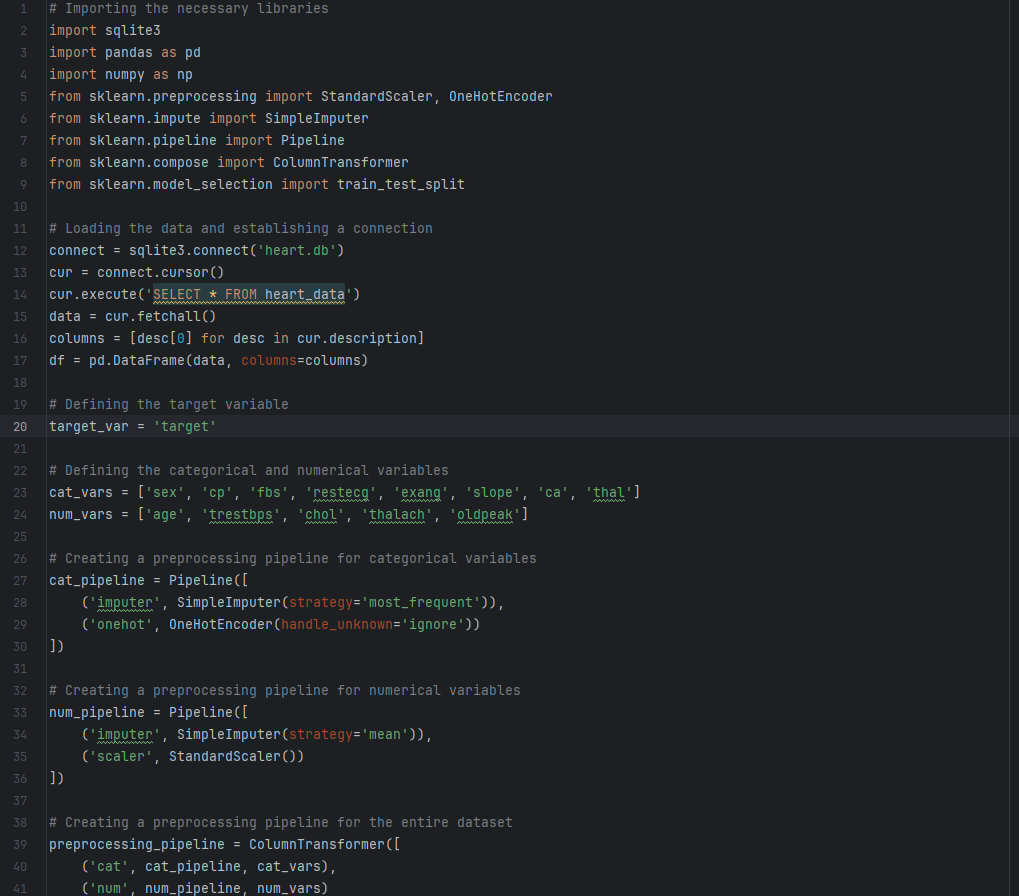
4. Thalac Distribution

- The distribution of thalach (max heart rate achieved) for class 0 is centered around 150-160, while class 1 is centered around 120-130, suggesting that people with heart diseases tend to have lower max heart rates.

5. Oldpeak Distribution

- The distribution of oldpeak (ST depression induced by exercise relative to rest) for class 0 is centered around 0-1, while class 1 is centred around 1-2, suggesting that people with heart diseases tend to have higher ST depression induced by exercise relative rest.

**Question 3**

3.1 Get your data ready for fitting a machine learning model on it by performing the appropriate preprocessing techniques.A screenshot of a computer program

Description automatically generated (87 Getting Your Data Ready Convert Data To Numbers | Scikit-learn Creating Machine Learning Models, 2021)

3.2. Select 3 appropriate machine learning models for your heart disease prediction problem. Provide a short explanation of each chosen model as well as two advantages and disadvantages of each. Use the three models to fit your data and perform predictions on it, then determine which model performs the best. Save the model to disk. Remember, this saved model will then be used to model your decision support system.

Machine learning models is a program that can find patterns or make decisions from a previously unseen dataset (databricks, 2024).

The following 3 models where used:

1. Logistic Regression

- Logistic regression is a linear model that predicts the probability of a binary response based on a set of predicor variables (ScienceDirect, 2014).

Advantages:

* Easy to interpret and implement (GeeksforGeeks, 2023).
* Fast training and prediction times.

Disadvantages

* Assumes a linear relationship between predictors and response (GeeksforGeeks, 2023).
* Can be sensitive to outliers and multicollinearity.

2. Random Forest

- Is an ensemble learning method that combines multiple decision trees to improve the accuracy and robustness of predictions (IBM, 2024).

Advantages

* Can handle large datasets and complex interactions between predictors.
* Robust to outliers and noisy data (geeksforgeeks, 2024).

Disadvantages

* Can be computationally expensive to train (geeksforgeeks, 2024).
* Difficult to interpret individual tree contributions.

3. Support Vector Machine (SVM)

- SVM is a linear or nonlinear model that finds the hyperplane that maximally separates the classes in the feature space (IBM, 2024).

Advantages

* Can handle high-dimensional data and nonlinear relationships (GeeksForGeeks, 2023).
* Robust to outliers and noisy data.

Disadvantages

* Can be computationally expensive to train.
* Requires careful tuning of hyperparameters (GeeksForGeeks, 2023).

A screenshot of a computer program

Description automatically generated A screen shot of a computer program

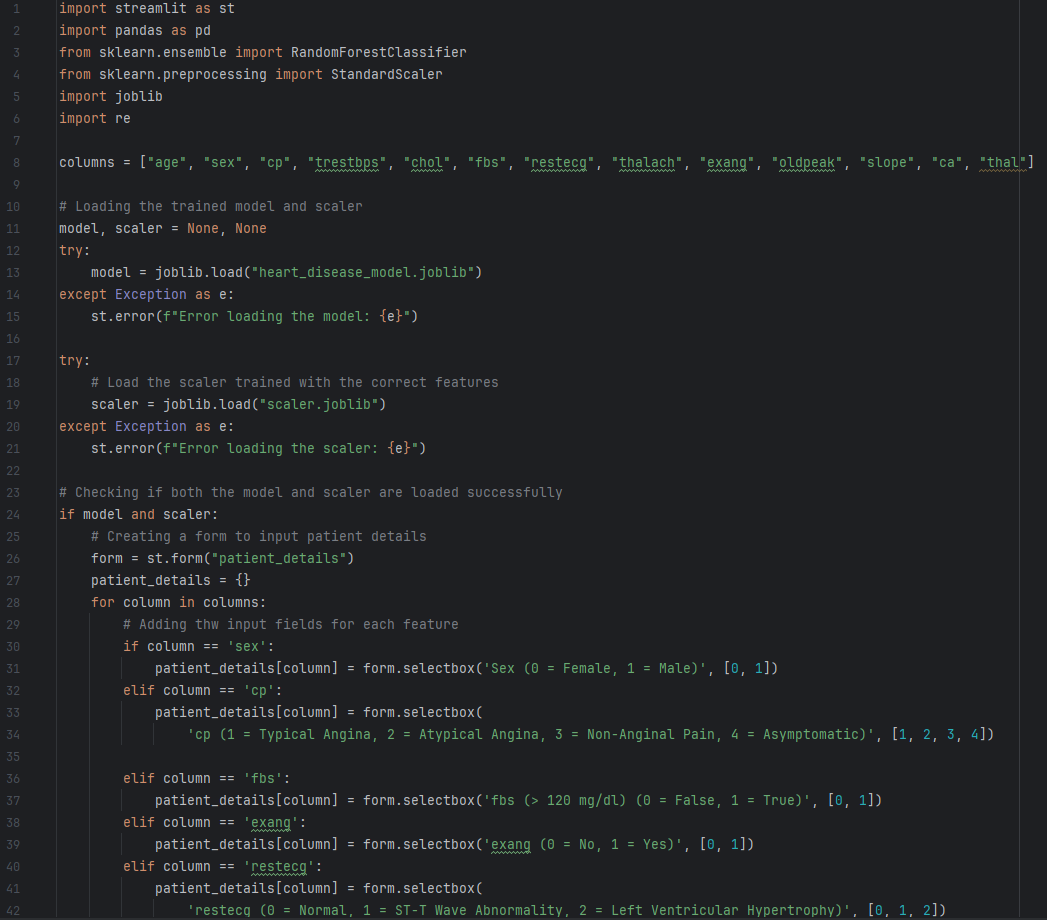
Description automatically generated A screen shot of a computer program

Description automatically generatedA screenshot of a computer

Description automatically generated

**Question 4**

Create a web application in streamlit that allows doctors to enter the details of patients (fill in the columnschosen for the model) to determine whether the patient likely suffers from heart disease so they may decideto send the patient for further tests or treatment. Add the link to your deployed app and github repo to the pdfsubmission doc. You should take the following into account when creating your application

 A screenshot of a computer program

Description automatically generated A screenshot of a computer program

Description automatically generated A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated