



Eduvos
Your Education. Your Future.

Eduvos (Pty) Ltd (formerly Pearson Institute of Higher Education) is registered with the Department of Higher Education and Training as a private higher education institution under the Higher Education Act, 101, of 1997. Registration Certificate number: 2001/HE07/008.

Date: Wednesday, 21 February 2024, 5:30 PM

ITDAA4-12 Assessments (2024)

Project

1. Project

Faculty:	Information Technology
Module Code:	ITDAA4-12
Module Name:	Data mining and Data administration
Content Writer:	Ms Taryn Michael
Internal Moderation:	Community of Practice
Copy Editor:	Mr Kyle Keens
Total Marks:	100
Submission Week:	Second Block week 6

This module is presented on NQF level 8.

5% will be deducted from the student's assignment mark for each calendar day the assignment is submitted late, up to a maximum of three calendar days. The penalty will be based on the official campus submission date.

Assignments submitted later than three calendar days after the deadline or not submitted will get 0%. [1]

This is an individual project

This project contributes 40% towards the final mark.

[1] Under no circumstances will assignments be accepted for marking after the assignments of other students have been marked and returned to the students.

2. Instructions to Students

1. Please ensure that your answer file (where applicable) is named as follows before submission: **Module Code – Assessment Type – Campus Name – Student Number**.
2. **You may submit one pdf containing your code snippets, outputs and any necessary explanations of the outputs.**
3. Remember to keep a copy of all submitted assignments.
4. All work must be typed.
5. Please note that you will be evaluated on your writing skills in all your assignments.
6. All work must be submitted through Turnitin. The full originality report will be automatically generated and available for the lecturer to assess. Negative marking will be applied if you are found guilty of plagiarism, poor writing skills, or if you have applied incorrect or insufficient referencing. (See the "instructions to students" book activity before this activity where the application of negative marking is explained.)
7. You are not allowed to offer your work for sale or to purchase the work of other students. This includes the use of professional assignment writers and websites, such as Essay Box. You are also not allowed to make use of artificial intelligence tools, such as ChatGPT, to create content and submit it as your own work. If this should happen, Eduvos reserves the right not to accept future submissions from you.

3. Project Overview

Project Objective:

The primary objective of this project is to create a decision support system as a web-based application, designed to assist healthcare professionals in addressing the challenge of diagnosing heart disease. The project is focused on accurately diagnosing individuals with heart disease.

You are thus provided with a csv file containing information about patients relating to their health, with a column indicating whether they suffer from heart disease or not. You need to use this dataset and create a decision support system to assist healthcare professionals in determining how likely a patient is to have heart disease, and thus whether they need to perform further tests to confirm this and treat the patient.

Deliverables:

Your deliverables for this project include:

- A pdf file with answers to questions 1 - 3.
- A deployed web application (DSS) using Streamlit

Format of questions in pdf doc:

Your answers in the pdf file will take on the following format:

- Code used to answer the question
- Code output
- Any explanations of the output as needed.

Format this as neatly as possible.

For Question 4, you will be required to deploy your machine learning model in a web application using Streamlit in Python. You can simply provide the **link to the deployed application** in your pdf doc (here is a link detailing how to deploy the app <https://docs.streamlit.io/streamlit-community-cloud/deploy-your-app>)

as well as the link to your **GitHub repo** containing the code for your Streamlit Application.

4. Dataset Overview

The dataset used in this project is a subset of the Heart Disease dataset from the [Original Heart Disease data repo](#).

It contains 76 attributes. However, the subset we are using only contains 14, and they are as follows:

Attribute	Description	Type
age	Patient's age in years	numerical
sex	sex (1 = male; 0 = female)	nominal/binomial
cp	chest pain type (1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic)	nominal
trestps	resting blood pressure (in mm Hg on admission to the hospital)	numerical
chol	serum cholestoral in mg/dl	numerical
fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)	binomial/nominal
restecg	resting electrocardiographic results (normal; abnormal; ventricular hypertrophy)	nominal
thalach	maximum heart rate achieved	numerical
exang	exercise induced angina (1 = yes; 0 = no)	binomial/nominal
oldpeak	ST depression induced by exercise relative to rest	numerical
slope	the slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping)	nominal
ca	number of major vessels colored by fluoroscopy (0 = mild; 1 = moderate; 3 = severe)	nominal
thal	Status of the heart (1 = normal; 2 = fixed defect; 3 = reversible defect)	nominal
target	(1 = heart disease; 0 = healthy)	binomial/nominal

Our field of interest is the "target" variable, which indicates whether the patient likely suffers from heart disease or not.

The objective of this project is to create a web application acting as a decision support system using data exploration and machine learning techniques to determine whether a patient has heart disease and thus needs to be treated or not.

5. Section A

Section A

Learning Objective

The objective of this project is to gain a practical understanding of the use and implementation of decision support systems in an organizational context. This project also aims to test the Python programming and problem solving skills of the student.

Project Topic

Decision Support System determining whether Patients have Heart Disease for Healthcare Professionals

Scope

Block 1 & 2, Weeks 1 - 7

Technical Aspects

- Include page numbers.
 - Format your questions as:
 - 1. Code snippet
 - 2. Code output
 - 3. Explanation of code output or any other necessary explanations.
 - Include all relevant diagrams.
-
- Ensure any diagrams, screenshots and PowerPoint presentations fit correctly on the page and are referenced.
 - Include a table of contents.
 - Use accurate Harvard referencing throughout the assignment.
 - Include a bibliography based on the applicable referencing method at the end of the assignment.
 - Include the completed Assignment/Project Coversheet (available on myLMS).
 - Run the assignment through the Turnitin software.

Plagiarism

Plagiarism is a very serious matter and if any part of your assignment is found to be plagiarised, you risk a zero grade.

5.1. Question 1

Question 1

10 Marks

Study the scenario and complete the question that follows:

Creating sqlite3 Database connection

The first steps in any project is to load, preprocessing and examine your data.

In this project you are given a csv file, however, typically in real-life scenarios you need to connect to a SQL database and perform queries on there. Thus in this section you will be uploading the csv file as a database table and connecting to it in python using the sqlite3 library.

Connecting to a SQLite database.

Create and set up a connection to a SQLite database that you will be reading your data from.

(10 Marks)

End of Question 1

5.2. Question 2

Question 2

30 Marks

Study the scenario and complete the questions that follow:

Data Preprocessing and Visualization

Once you have established a connection to the database, you need to transform (preprocess) your data to get it into a more consistent, accurate, and reliable format. This will then allow you to explore and visualize the data to gain significant insights into it.

2.1 Preprocessing and visualizing the data

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

(4 Marks)

- 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.

(16 marks)

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

(10 marks)

End of Question 2

5.3. Question 3

Question 3

30 Marks

Study the scenario and complete the questions that follow:

Modelling heart disease prediction problem through machine learning

After getting a better understanding of your data, you are not able to begin the modelling process. You will use machine learning techniques to predict whether a patient suffers from heart disease or not.

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

(10 Marks)

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.

End of Question 3

5.4. Question 4

Question 4

30 Marks

Study the scenario and complete the question that follows:

Web application using streamlit

Now that you have created and saved your model, you can deploy your model in a web application using streamlit for medical practitioners to use.

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

- The functionality of your application
- The usability
- The design of the application
- The code design/structure

- Your code and application documentation
- Error-handling

(30 Marks)

End of Question 4