Faculty of Computer Science & Information Technology

PRACTICAL OF AI

(WID3014)

Assignment 2- 15%

DUE DATE: 20th DECEMBER 2023

STUDENT DECLARATION

1. *I confirm that I am aware of the University's Regulation Governing Cheating in a University Test and Assignment and of the guidance issued by the Faculty of Computer Science & Information Technology concerning plagiarism and proper academic practice and that the assessed work now submitted is in accordance with this regulationand guidance.*
2. *I understand that, unless already agreed with the Faculty of Computer Science & Information Technology, assessed work may not be submitted that has previously been submitted, either in whole or in part, at this or any other institution.*
3. *I recognise that should evidence emerge that my work fails to comply with either of the above declarations, then I may be liable to proceedings under Regulation.*

No	Student Name	Student ID	Date	Score (15%)
1				
2				

Notes:

1) Course Learning Outcome: On completion of this alternative assessment, students should be able to:

   CLO1: Identify solution approach that is suitable for the stated problem

2) The tool for doing all the tasks is **"Python"**.
3) If you use any references (articles, web pages, etc.), you should cite the references in your writing and add the references at the bottom of your report.
4) Using AI tools is not recommended. But, in the case of using any AI tools, below instructions must be followed:
   a. The AI tool must be cited properly.
   b. The output of the AI tool must be interpreted.
   c. At least 3 improvements to the AI-suggested answers must be discussed.
   d. Student needs to present the report in the class, defend the answers, and provide required justifications if needed.
5) Copying, cheating, attempts to cheat, plagiarism, collusion, and any other attempts to gain an unfair advantage in assessment result in being awarded 0 marks to all parties concerned.
6) All the submitted documents will be cross-checked with other students' during your presentation.
7) Severe disciplinary action will be taken against those caught violating assessment rules such as colluding, plagiarizing or transcribing.
8) The assignment submission document should be 10 - 20 pages in total with a spacing of 1.5 and a font of 12pt Times New Roman.

Submission Requirements

- ✓ A Report of your answers and write-ups (in PDF) including
  - ➢ Cover page
  - ➢ References (if any)
- ✓ Python code files (.ipynb)

Case Study Breast Cancer:

Breast cancer is the most common cancer amongst women in the world. It accounts for 25% of all cancer cases, and affected over 2.1 Million people in 2015 alone. It starts when cells in the breast begin to grow out of control. These cells usually form tumors that can be seen via X-ray or felt as lumps in the breast area. Breast cancer prediction using machine learning is an active area of research and application that aims to develop models capable of analyzing medical data to predict the likelihood of breast cancer or to assist medical professionals in diagnosing the disease. Machine learning techniques can help analyze complex patterns in large datasets and provide insights that aid in early detection and improved patient outcomes.

The key challenges against it's detection is how to classify tumors into malignant (cancerous) or benign (non cancerous). We ask you to complete the analysis of classifying these tumors using machine learning with the given dataset.

Objective:

- Understand the Dataset & cleanup (if required).
- Build classification models to predict whether the cancer type is Malignant or Benign.
- Also fine-tune the hyperparameters & compare the evaluation metrics of various classification algorithms.

Your task is to make a report based on the following actions:

	Task items	Score
1	1.1. Import the dataset and explore the data. 1.2. Provide some statistics base on the given dataset. Indicate missing, duplications, and extreme/unrealistic values if there are any.	2 marks
2	2.1. Clean the dataset from missing values, duplications, and extreme/unrealistic values. Give why you choose to remove the datum instead of replacing it or vice versa. 2.2. Attach the clean dataset to the final report.	2 marks
3	3.1. Create a correlation matrix. Interpret the correlation coefficients as displayed on the matrix. The interpretation is to answer the following question: 3.2. What correlations exist? 3.3. How strong are they? **Note:** Interpret at least a pair of attributes	3 marks
4	4.1. Convert the categorical attribute to numerical values 4.2. Split dataset to 80% training and 20% testing data. What theory you are using for splitting the dataset. 4.3. Create a model using A (ONE) classification algorithm of your choice. 4.4. Visualize the model outputs, add the figures to your report, and describe what each figure illustrates.	4 marks
5	5.1. Evaluate the performance of the model using at least 2 metrics. 5.2. Describe each metric and explain why you chose it. 5.3. Visualize the outputs on examining the test dataset, add the figures into your report, and describe what each figure illustrates.	3 marks
6	6.1. Were any attributes dropped from the data set as non-predictors? If so,which ones and why do you think they weren't effective predictors?	1 mark

Marking Rubric Assignment 2_15%

Criteria	Score (Percentage of the allocated marks for each task)			
	Excellent	**Good**	**Average**	**Poor**
	>= 13%	< 13% >= 10%	< 10% >= 5%	< 5%
Q1: Data Exploration	All the presented statistics and types of data and dataset(s) are identified correctly. All types of noises or errors are found in those data. The answer is supported with some samples/records from the dataset(s). The writing is clear and understandable. The similarity is less than 2%.	Most of the presented statistics and types of data and dataset(s) are identified correctly. Most of the types of noises or errors are found in those data. The answer is supported with some samples/records from the dataset(s). The writing is clear and understandable. The similarity is less than 4%.	Some of the presented statistics and types of data and dataset(s) are identified correctly. Some types of noises or errors are found in those data. The answer is not supported with samples/records from the dataset(s). The writing is understandable. The similarity is less than 5%.	The presented statistics and the types of data and dataset(s) are not identified correctly. AND/OR some types of noises or errors of the dataset are not highlighted. The answer is not supported with samples/records from the dataset(s). The writing is not clear and understandable. The similarity is more than 5%.
Q2: Preprocessing concept, and knowledge	All the preprocessing techniques are identified correctly and explained clearly. The answer is supported with evidence and references. The writing is clear and understandable. The similarity is less than 2%.	Most of the preprocessing techniques are identified correctly and explained clearly. The answer is supported with evidence and references. The writing is clear and understandable. The similarity is less than 4%.	Some of the preprocessing techniques are identified and explained acceptedly. The answer is supported with evidence and references. The writing is clear and understandable. The similarity is less than 5%.	Most of the preprocessing techniques are not identified and explained. The answer is not supported with evidence and references. The writing is not clear and understandable. The similarity is more than 5%.
Q3: Correlation Matrix	The correlation Matrix is generated correctly. The answers are valid and acceptable. The program/code is correct. The justification is accurate, clear, and valid. The similarity is less than 2%.	The correlation Matrix is generated correctly. The answers are valid and acceptable. The program/code is correct. The justification is acceptable but not well written. The similarity is less than 4%.	The correlation Matrix is generated with minor mistakes. The answers are valid with minor mistakes. The program/code is correct. The justification is acceptable but not well written. The similarity is less than 5%.	The correlation Matrix is generated with major mistakes. The answers are not valid and have major mistakes. The program/code is not correct. The justification is not acceptable and is not well written. The similarity is more than 5%.
Q4: Data modelling and development	The answer is valid and acceptable. The program/code is correct. The justification is accurate, clear, and valid. The similarity is less than 2%.	The answer is valid and acceptable. The program/code is correct. The justification is acceptable but not well written. The similarity is less than 4%.	The answer is valid and acceptable. The program/code is correct. But the justification is not well written. The similarity is less than 5%.	The answer is not valid and acceptable. The program/code is not correct. The justification is not accurate, not clear, and not valid. The similarity is more than 5%.
Q5: evaluation and Visualizing the performance of the model	The answer is valid and acceptable. The program/code is correct. The justification is accurate, clear, and valid. The similarity is less than 2%.	The answer is valid and acceptable. The program/code is correct. The justification is acceptable but not well written. The similarity is less than 4%.	The answer is valid and acceptable. The program/code is correct. But the justification is not well written. The similarity is less than 5%.	The answer is not valid and acceptable. The program/code is not correct. The justification is not accurate, not clear, and not valid. The similarity is more than 5%.
Q6: data analytics knowledge	The answer is correct.All the provided evidences are relevant and correct.The writing is clear and understandable. The similarity is less than 2%.	The answer is correct. Most of the provided evidences are relevant and correct the writingis clear and understandable. The similarity is less than 4%.	The answer is correct. Some of the provided evidences are relevant and correct the writingis clear and understandable. The similarity is less than 5%.	The answer is correct. A few of the provided evidences are relevant and correct the writingis not clear and understandable. The similarity is more than 5%.

-End-