ASSESSMENT 2 – 18 May 2024

SUBJECT & LEVEL	**:**	**TECHNICAL PROGRAMMING 2**
SUBJECT CODE	**:**	**TPRO200/TLPR200**
DEPARTMENT	**:**	**INFORMATION AND COMMUNICATION TECHNOLOGY (ICT)**
QUALIFICATION	**:**	**DIPLOMA: INFORMATION TECHNOLOGY (IT)**
EXAMINER (S)	**:**	**MR X PIYOSE**
MODERATOR (S)	**:**	**DR MB MUTANGA**
FULL MARKS	**:**	**110**
TOTAL MARKS	**:**	**100**
NUMBER OF PAGES (incl. cover page)	**:**	**12**

DURATION OF ASSESSMENT	**:**	**240 MINUTES (4 hours)**
VENUE OF ASSESSMENT	**:**	**D9/D10 and D2**
TIME OF ASSESSMENT	**:**	**9h00 TO 13h00**
REQUIREMENTS	**:**	**- REFER TO THE NEXT PAGE FOR MANDOTORY REQUIREMENTS**

REQUIREMENTS AND IMPORTANT INFORMATION (Note: students will receive these requirements and READ them 48 hours before the test is written):

- VERY IMPORTANT: This test MUST be written on **jupyter notebook** NOT Visual Studio code. 40% will be deducted from the final mark if this instruction is not followed.
- **Dataset**: The students are expected to find a dataset that is related to the problem statement. Once you have found an informal settlement or human settlement dataset – you must use it (don't over analyse or over think it). Please watch the following YouTubeVideo: https://www.youtube.com/watch?v=PExdWWcxmro. More than one students can use the same dataset for this test, however those students answers will be strictly monitored by the assessor. Also note that although the problem statement is South African based, your data set does not have to be south African or African for that matter.
- Python version 3xxxxx should be installed. Note that python version 2xxx and version 1xxx are not permitted.
- All the following packages should have been installed already as per Activities 1, 2, 3 4, 5, 6, 7, 8 and 9 students were required to do. The packages are: **numpy, pandas, scipy**, **scikit-learn**, **matplotlib** and **seaborn**.
- This test constitutes 75% of your weight 2 ITS mark, the other 25% will be taken from Practical Activity 5, 6, 7, 8 and 9.
- As per the norm in the subject, GIT should have been installed on student laptop for effective source code management.
- Each student should have a GitHub account.
- Very important: students that are working on Git bash or command line interface from a windows laptop should NEVER use the **cls** command during the test; and students that are working on a terminal from a Mac laptop should NEVER use the **clear** command.
- Your lecturers GitHub username is **xpiyose,** or if you prefer to use the email address, it is: **xpiyose@gmail.com.**
- You will required to answer ALL questions.
- Negative marking may be used for syntax errors.
- Ensure your code is clean, well-commented, and adheres to Python best practices.

- **VERY VERY IMPORTANT:** if your test duration is over before you finish the test, you must stand and wait for your lecturer or invigilator to give you further instructions. Students will be expected to sign a mandatory checklist that will be distributed during the test, before they leave the venue. When you are done with your test, DO NOT LEAVE the venue, simply raise your hand and inform the invigilator that you are done!!!

Theme: Human settlement for sustainable development

Problem statement:

Informal settlements, also known as "squatter camps," "skwatta camps," "mijondolo," or "mikhukhu," present significant challenges regarding human settlement in the Republic of South Africa. According to Statistics South Africa (Stats SA), there are approximately eight new informal settlements emerging every six months across the country. Typically, these settlements initially accommodate around 100 families. These communities often lack essential infrastructure, including sanitation, clean water, and electricity. The absence of such basic services exacerbates health issues, heightens social inequality, and limits economic opportunities for residents.

The South African government, in conjunction with various non-governmental organizations (NGOs), has undertaken efforts to address these challenges. However, the complexity and scale of the problem necessitate innovative, data-driven solutions.

Your task is to develop a system that can identify trends related to the emergence of new informal settlements. This system should analyze data pertinent to informal settlements and propose solutions aimed at improving living conditions and supporting sustainable development. By leveraging technology and data analysis, your goal is to create a comprehensive approach to mitigate the issues associated with informal settlements and promote equitable and sustainable growth.

Generic Requirements:

You are an artificial intelligence architect/developer working for a non-profit organization focused on improving living conditions in informal settlements across South Africa. Your team has been tasked with identifying key issues in these settlements and proposing feasible solutions. You need to source relevant data, clean and analyze it, and use algorithms to uncover insights that can guide your recommendations. You MUST use the rubric as a guide on what you are required to do.

Specific requirements (SR):

- **SR1 - Data Sourcing (5 marks):**

Find a publicly available dataset related to informal settlements. Document where the dataset was found and justify its suitability for the analysis. You should consider factors like data completeness, relevance, and the credibility of the source. Your justification should be written on a markdown cell of your jupyter notebook. Hint: study the columns of the datasets to see if they will be relevant to the theme and choose at least 2 columns to justify the relevancy of this dataset to the theme. Your dataset must be on a **.csv** format.

- **SR2 – Pre-processing (15 marks):**

Write Python code to understand and clean the dataset. This should include:

  - Understanding the size, data types, columns, rows, data slicing and indexing of the dataset as whole.
  - identify missing data
  - Normalizing or standardizing data if required
  - Describe the data types that are used in relation to machine learning by using one column for your data set for each machine learning or statistical dataset on a markdown cell.

- **SR 3 - Exploratory Data Analysis (EDA) (35 marks):**

Write a python code that will perform an EDA to identify trends and key insights into the conditions and demographics of informal settlements. Present your findings using at least:

  - Charts and graphs: you must use any three different basic infographic charts for your analysis and any two advanced charts that you have learned in this subject.
  - Summary statistics (e.g., mean, median, standard deviation)

Hint: in some charts highlight significant patterns, such as areas with the highest population density, regions most lacking in basic services, and any correlations between variables (depending on your dataset columns).

- **SR 4.1: Algorithm Implementation: Sort algorithm (30 marks):**

For optimization and performance purpose, implement any sort algorithm between bubble sort, insertion sort and selection to sort any column in your dataset in either ascending or descending order (your choice). Use only ONE sort algorithm from the ones mentioned.

An algorithm in machine learning at times goes to what is called a decision tree in order to reach it's final decision. The final decision is usually based on identifying parent nodes, child nodes and leaf nodes. A heap sorting algorithm do provide basics of decision tree algorithm. Assume that in your data set you have a column called age, which records the ages of the people that are staying at each "mkhukhu". Assume that for one of the households ages are 6 9 8 2 45 63 81. The ages 45 63 and 81 are already sorted. Using pen and paper, heapify the rest of the ages such that they are all sorted in ascending order. You must provide clear reasons for each iteration. You must take a screenshot of your iterations and save them as heap1.jpeg, heap2.jpeg, etc.

- **SR 4.2: Algorithm Implementation: Search algorithm (20 marks):**

Machine learning algorithms use inputs from the user in order to search for the best answer to provide to the user. Many well known search engines such as Google, bing, and AI information generators such as chatgpt, jenni ai, etc uses search algorithms in order to provide the best answer to the user. In this subject you have learned 2 searching algorithms namely linear and binary search.

  - Explain which of these algorithms performs faster and provide reasons using a markdown cell in your jupyter notebook. (6)
  - Write a python code that will search for a specific element in your dataset using both algorithms. (14)

- **SR 5: Complexity Analysis: (5 marks):**

Choose one algorithm that you wrote in SR 4.1 or SR 4.2 to explain the time and space complexity of your implemented algorithms. Explain the efficiency of the algorithms in terms of their Big-O notation. You must use a markdown cell for this question.

Submission requirements:

- Push your files, i.e jupyter notebooks (you must create a jupyter notebook for each SR) as well as a copy of your screenshots as per SR 4.1 into your github account.
- Add your lecturer as a collaborator.

---------- end of the test----------

EVALUATION CRITERIA			
STUDENT NUMBER:			
SR1 – DATA SOURCE	**Mark**	**Examiner**	**Moderator**
The dataset selected is pertinent to the specified theme. A markdown cell within the Jupyter notebook provides a comprehensive justification for the dataset's selection, focusing on its relevance, completeness, and the credibility of its source. Additionally, the markdown cell includes a hyperlink to the dataset's source for reference.	(5)		
SR 2 – DATA UNDERSTANDING AND PRE-PROCESSING TECHNIQUES	**-**	**-**	**-**
The Python code used provides a comprehensive understanding of the dataset, including its size, data types, the number of columns and rows, and how to slice and index the data effectively. *This foundational understanding is crucial for any further data analysis or processing steps and will guide the assessor on how to mark the rest of your answers for this test.*	(8)		
The code must identify and handle any missing data within the dataset. This involves detecting null or missing values. Comment on appropriate methods or ways to ensure the dataset integrity and accuracy. For instance, do you	(3)		

have any missing values? will it be appropriate to train the machine learning model with missing values from the dataset?			
A detailed description of the data types used in the dataset should be provided, focusing on their relevance to machine learning or statistical analysis. This description should be written in a markdown cell and should include examples of how each data type is utilized in one column of the dataset, highlighting their importance for the intended analysis.	(4)		
SR3 – EDA	**-**	**-**	**-**
Python code provided creates and presents findings using at least three different basic infographic charts (e.g., bar charts, line graphs, pie charts) and two advanced charts (e.g., heatmaps, scatter plots with regression lines) learned in this subject. These visualizations should effectively communicate key insights and trends within the dataset.	(25)		
Summary statistics for the dataset, including measures such as the mean, median, and standard deviation have been computed. These statistics should give a comprehensive overview of the data's central tendencies and variability.	(5)		
Significant patterns and insights have been highlighted. For instance, identify and mark areas with the highest population density, regions most lacking in	(5)		

basic services like sanitation and clean water, and any notable correlations between different variables (e.g., the relationship between income levels and access to electricity). These patterns should be clearly indicated to provide a deeper understanding of the data.			
SR4.1 ALGORITHM IMPLEMENTATION: SORT	**-**	**-**	**-**
Any of the following sort algorithms (bubble sort, insertion sort, or selection sort) has been implemented to sort a dataset column either ascending or descending order. The implementation demonstrates a clear understanding of the chosen algorithm's mechanics, ensuring that the code is efficient, runs without errors, and handles edge cases gracefully. Hint: The sorting process must be thoroughly documented, with explanations for each step of the algorithm and the rationale behind selecting the sorting order. Additionally, the code should include comments and explanations that enhance readability and comprehension, making it clear how the algorithm operates and why specific steps were taken.	(10)		
The heapification process is clear, demonstrating the understanding of parent, child and leaf nodes as used for decision tree algorithm and heap data structure in machine learning. The iterations and their corresponding explanations are well documented and clear into how the algorithm got to its final decision.	(20)		

SR4.2 ALGORITHM IMPLEMENTATION: SEARCHING	-	-	-
In a markdown cell, a detailed explanation which search algorithm, between linear search and binary search, performs faster. A detailed analysis of the time complexities of both algorithms, highlighting which algorithm performs faster from the two for big O notation O(log n) compared to O(n). Reasons for this performance difference are clear.	(6)		
A Python code to search for a specific element in a dataset using both linear search and binary search algorithms exist, and it works. The implementation demonstrates a clear understanding of how each algorithm works. A clear and well-commented code for both algorithms has been provided, ensuring that the search processes are correctly implemented and easy to follow. This will demonstrate your ability to apply theoretical knowledge of search algorithms to practical problems. Hint: students should remember that AI generators, search engines, maps and so on are using a search algorithm to provide accurate feedback to the user.	(14)		
SR 5 – Complexity analysis	-	-	-
Time and space complexity has been explained using any algorithm that the student created for the test.	(6)		
Total	100		