MODULE NAME:	MODULE CODE:
PROGRAMMING FOR DATA ANALYTICS 1	PDAN8411/w

ASSESSMENT TYPE: POE (PAPER)

TOTAL MARK ALLOCATION: 300 MARKS

TOTAL HOURS: A minimum of 30 HOURS is suggested to complete this assessment

By submitting this assignment, you acknowledge that you have read and understood all the rules as per the terms in the registration contract, in particular the assignment and assessment rules in The IIE Assessment Strategy and Policy (IIE009), the intellectual integrity and plagiarism rules in the Intellectual Integrity Policy (IIE023), as well as any rules and regulations published in the student portal.

INSTRUCTIONS:

1. ***No material may be copied from original sources, even if referenced correctly, unless it is a direct quote indicated with quotation marks. No more than 10% of the assignment may consist of direct quotes.***
2. ***Make a copy of your assignment before handing it in.***
3. *Assignments must be typed unless otherwise specified.*
4. *Begin each section on a new page.*
5. *Follow all instructions on the PoE cover sheet.*
6. *This is an individual assignment.*

Referencing Rubric

Providing evidence based on valid and referenced academic sources is a fundamental educational principle and the cornerstone of high-quality academic work. Hence, The IIE considers it essential to develop the referencing skills of our students in our commitment to achieve high academic standards. Part of achieving these high standards is referencing in a way that is consistent, technically correct and congruent. This is not plagiarism, which is handled differently.

Poor quality formatting in your referencing will result in a penalty **<u>of a maximum of ten percent</u> being deducted from the percentage awarded**, according to the following guidelines. Please note, however, that **evidence of plagiarism in the form of copied or uncited work (not referenced), absent reference lists, or exceptionally poor referencing, may result in action being taken in accordance with The IIE's Intellectual Integrity Policy (0023)**.

Markers are required to provide feedback to students by indicating **<u>(circling/underlining) the information that best describes the student's work.</u>**

<u>Minor technical referencing errors: 5% deduction from the overall percentage</u> – the student's work contains **<u>five or more errors</u>** listed in the minor errors column in the table below.

<u>Major technical referencing errors: 10% deduction from the overall percentage</u> – the student's work contains **<u>five or more errors</u>** listed in the major errors column in the table below**.**

<u>If both minor and major errors</u> are indicated, then 10% only (and not 5% or 15%) is deducted from the overall percentage. The examples provided below are not exhaustive but are provided to illustrate the error

<u>Required:</u> **Technically correct referencing style**	**<u>Minor errors</u> in technical correctness of referencing style** **Deduct 5% from percentage awarded**	**<u>Major errors</u> in technical correctness of referencing style** **Deduct 10% from percentage awarded**
<u>Consistency</u> • The same referencing format has been used for all in-text references and in the bibliography/reference list.	Minor inconsistencies. • The referencing style is generally consistent, but there are one or two changes in the format of in-text referencing and/or in the bibliography. • For example, page numbers for direct quotes (in-text) have been provided for one source, but not in another instance. Two book chapters (bibliography) have been referenced in the bibliography in two different formats.	Major inconsistencies. • Poor and inconsistent referencing style used in-text and/or in the bibliography/ reference list. • Multiple formats for the same type of referencing have been used. • For example, the format for direct quotes (in-text) and/or book chapters (bibliography/ reference list) is different across multiple instances.
<u>Technical correctness</u> Referencing format is technically correct throughout the submission. Position of the reference: a reference is directly associated with every concept or idea. For example, quotation marks, page numbers, years, etc. are applied correctly, sources in the bibliography/reference list are correctly presented.	**Generally, technically correct with some minor errors.** • The correct referencing format has been consistently used, but there are one or two errors. • Concepts and ideas are typically referenced, but a reference is missing from one small section of the work. • Position of the references: references are only given at the beginning or end of every paragraph. • For example, the student has incorrectly presented direct quotes (in-text) and/or book chapters (bibliography/reference list).	**Technically incorrect.** • The referencing format is incorrect. • Concepts and ideas are typically referenced, but a reference is missing from small sections of the work. • Position of the references: references are only given at the beginning or end of large sections of work. • For example, incorrect author information is provided, no year of publication is provided, quotation marks and/or page numbers for direct quotes missing, page numbers are provided for paraphrased material, the incorrect punctuation is used (in-text); the bibliography/reference list is not in alphabetical order, the incorrect format for a book chapter/journal article is used, information is missing e.g. no place of publication had been provided (bibliography); repeated sources on the reference list.
Congruence between in-text referencing and bibliography/ reference list • All sources are accurately reflected and are all accurately included in the bibliography/ reference list.	**Generally, congruence between the in-text referencing and the bibliography/ reference list with one or two errors.** • There is largely a match between the sources presented in-text and the bibliography. • For example, a source appears in the text, but not in the bibliography/ reference list or vice versa.	**A lack of congruence between the in-text referencing and the bibliography.** • No relationship/several incongruencies between the in-text referencing and the bibliography/reference list. • For example, sources are included in-text, but not in the bibliography and vice versa, a link, rather than the actual reference is provided in the bibliography.
In summary: the recording of references is accurate and complete.	In summary, at least **80%** of the sources are correctly reflected and included in a reference list.	In summary, at least **60%** of the sources are incorrectly reflected and/or not included in reference list.

Overall Feedback about the consistency, technical correctness and congruence between in-text referencing and bibliography:

....

....

Background

This portfolio of evidence consists of three tasks. Task 1 focuses on linear regression, as task 2 is about classification and model improvement, and the final Portfolio of Evidence submission requires pipelines and text data. This portfolio of evidence will require additional research to complete.

Instructions :

For each task, you will be required to find an **open data set** that can be used for analysis. Include the data in your submission and remember that you must include a **reference** to where you found your data. You may choose a new data set for each task or use one of the data sets in your course material. Do choose wisely so you can analyse the data using the **required method** in each task. You can find data sets here: https://www.kaggle.com/. Please note that there is sample code provided on this website. You are welcome to refer to the sample code as long as you do not directly copy code and attribute the original sources.

You are required to submit your work in the format of a **Jupyter notebook** for each task. The Jupyter notebook must contain the **reference** to where you found the open data set, the **Python code** to perform the relevant analysis of the data, and an **explanation of the process** that you followed to perform the analysis. The Jupyter notebook must also contain any extra instructions that should be taken note of when running the code.

The specific requirements for each task follow below.

Tip: Read the rubric at the end of this document in detail to make sure that you meet all the requirements.

Task 1 — Linear Regression (Marks: 100)

At the end of this specific task, students should be able to:

- *Create a program to visualise data using graphs;*
- *Apply supervised learning algorithms to solve problems.*

Choose a data set that can be analysed using linear regression. You may refer to how we classified the Iris in chapter 1 to assist you.

Create a Jupyter notebook that includes:

- A reference to the data set source.
- An explanation of why the data set is appropriate for linear regression.
- An explanation of what analysis is going to be performed on the dataset. What is the question that the analysis will answer? Pay special attention to:
    - How you plan to verify your results
    - What features you need to extract
    - How you will fit your data and your considerations for over and under fitting
    - Which libraries you will import and use
    - Which predications you will make
    - How you will visualize your data.

- Python code to load and analyse the data and display results graphically, with explanations of the process.
- Any extra instructions necessary to successfully run the code.

Submit the following:

- Jupyter notebook
- Data set

Task 2 — Classification and Model Improvement (Marks: 100)

At the end of this specific task, students should be able to:

- *Create a program to visualise data using graphs;*
- *Apply supervised learning algorithms to solve problems;*
- *Apply evaluation metrics in model selection.*

Choose a data set that can be analysed using classification methods. In this task, it is also important to evaluate and improve on the model and document the whole process.

Create a Jupyter notebook that includes:

- A reference to the data set source.
- An explanation of why the data set is appropriate for classification.
- An explanation of what analysis is going to be performed on the dataset. What is the question that the analysis will answer? Pay special attention to:
    - Training and testing the Naïve Bayes Classifier
    - The accuracy of your predictions
    - How you plan to plot your results
    - Consider how changing the training /test splits will affect your predictions and accuracy
    - Compare the accuracy of the K-NN algorithm to the Naïve Bayes Classifier.
- Python code to load and analyse the data and display results graphically, with explanations of the process.
- An explanation of how the model was evaluated and improved.
- Any extra instructions necessary to successfully run the code.

Submit the following:

- Jupyter notebook
- Data set

Portfolio of Evidence — Pipelines and Text Data (Marks: 100)

At the end of this specific task, students should be able to:

- *Apply pipelines in Python to chain multiple steps;*
- *Apply the term frequency–inverse document frequency method to text data;*
- *Apply advanced tokenisation to text data;*
- *Apply stemming to text data;*
- *Apply lemmatisation to text data;*
- *Apply Latent Dirichlet Allocation to text data.*

Choose a data set containing text data, that can be analysed using the techniques mentioned above. Make use of a pipeline to sequence the steps required for the analysis. In this task, it is also important to evaluate and improve on the model and document the whole process.

Create a Jupyter notebook that includes:

- A reference to the data set source.
- An explanation of why the data set is appropriate for text processing.
- An explanation of what analysis is going to be performed on the dataset. What is the question that the analysis will answer?
- Python code to load and analyse the data and display results graphically, with explanations of the process. Note that a pipeline is a requirement here.
- An explanation of how the model was evaluated and improved.
- Any extra instructions necessary to successfully run the code.

Submit the following:

- Jupyter notebook
- Data set

Assessment Sheet (Marking Rubric)

Please note: Tear off this section and **attach** it to your work when you submit it/ If this is an online submission, then this information needs to be included in the online submission.

MODULE NAME:	**MODULE CODE:**
PROGRAMMING FOR DATA ANALYTICS 1	**PDAN8411**

STUDENT NAME:	
STUDENT NUMBER:	

Marking Criteria	***Fail/Does not meet the required standard (0% to 49%)***	***Average/meets the required standard (50% to 64%)***	***Above average/is above the required standard (65% to 74%)***	**Excellent** ***(75% - 100%)***	**Feedback**
		TASK 1 – LINEAR REGRESSION			
Knowledge: **Explanation of what linear regression is.** **[10 Marks]**	• No explanation included; • Explanation is unclear or incorrect.	• A correct but basic explanation is included.	• An explanation is included with additional details beyond the basics.	• A comprehensive explanation is included.	
	0 – 4 Marks	**5 – 6 Marks**	**7 Marks**	**8 – 10 Marks**	

Marking Criteria	*Fail/Does not meet the required standard (0% to 49%)*	*Average/meets the required standard (50% to 64%)*	*Above average/is above the required standard (65% to 74%)*	**Excellent** *(75% - 100%)*	**Feedback**
TASK 1 – LINEAR REGRESSION					
Application: **Explanation of why the chosen data set is appropriate for analysis with linear regression.** **[10 Marks]**	• No explanation is included; • The explanation does not link up with the theory at all.	• A basic explanation is included that links up with the theory; • An explanation is included that mostly lines up with the theory.	• An explanation is provided that clearly links up with the theory, but could have more detail.	• A comprehensive explanation is included that applies the theory.	
	0 – 4 Marks	**5 – 6 Marks**	**7 Marks**	**8 – 10 Marks**	
Application: **An explanation is included of what the analysis is that will be conducted on the data set.** **[10 Marks]**	• No explanation included; • The explanation lacks details in terms of what the analysis will accomplish.	• An explanation is included that provides the bare basics of what the analysis will accomplish.	• An explanation is included that mostly explains what the analysis will be about, but could be more detailed.	• A well-motivated explanation is included that clearly details what the purpose of the analysis is.	
	0 – 4 Marks	**5 – 6 Marks**	**7 Marks**	**8 – 10 Marks**	

Marking Criteria	*Fail/Does not meet the required standard (0% to 49%)*	*Average/meets the required standard (50% to 64%)*	*Above average/is above the required standard (65% to 74%)*	**Excellent** *(75% - 100%)*	**Feedback**
TASK 1 – LINEAR REGRESSION					
Application: **Python code is included that correctly performs the analysis.** **[40 Marks]**	• No Python code is included; • Python code does not run at all; • Python code only partially works.	• Python code mostly works correctly with one or two issues under normal usage scenarios.	• Python code works correctly under most usage scenarios.	• Python code works perfectly and will be able to handle any requests without issues.	
	0 – 19 Marks	**20 – 26 Marks**	**27 – 29 Marks**	**30 – 40 Marks**	
Application: **The analysis is represented well using graphical elements** **[20 Marks]**	• No graphical elements are included in the analysis; • Analysis is displayed but only by means of text output.	• Graphical elements are used but are missing elements that help the user of the output to understand the meaning of the display.	• Graphic elements are used in a way that is mostly clear to the user of the output.	• The graphical elements clearly display the results of the analysis in a way that is easy to interpret by the user of the output.	
	0 – 9 Marks	**10 – 13 Marks**	**14 Marks**	**15 – 20 Marks**	

Marking Criteria	*Fail/Does not meet the required standard (0% to 49%)*	*Average/meets the required standard (50% to 64%)*	*Above average/is above the required standard (65% to 74%)*	**Excellent** *(75% - 100%)*	**Feedback**
		TASK 1 – LINEAR REGRESSION			
Knowledge: **An explanation is included of the process used for the data analysis.** **[10 Marks]**	• No explanation is included; • The explanation is unclear; • The explanation is not relevant to the specific code.	• The explanation includes all the steps but lacks detail; • The explanation includes only some of the steps in detail.	• The explanation covers all the steps but lacks detail in some cases.	• The explanation is fully detailed, explaining every step of the process.	
	0 – 4 Marks	**5 – 6 Marks**	**7 Marks**	**8 – 10 Marks**	

Marking Criteria	***Fail/Does not meet the required standard (0% to 49%)***	***Average/meets the required standard (50% to 64%)***	***Above average/is above the required standard (65% to 74%)***	**Excellent** *(75% - 100%)*	**Feedback**
TASK 2 – CLASSIFICATION AND MODEL IMPROVEMENT					
Knowledge: **Explanation of what classification is.** **[10 Marks]**	• No explanation included; • Explanation is unclear or incorrect.	• A correct but basic explanation is included.	• An explanation is included with additional details beyond the basics.	• A comprehensive explanation is included.	
	0 – 4 Marks	**5 – 6 Marks**	**7 Marks**	**8 – 10 Marks**	
Application: **Explanation of why the chosen data set is appropriate for analysis with classification.** **[10 Marks]**	• No explanation is included; • The explanation does not link up with the theory at all.	• A basic explanation is included that links up with the theory; • An explanation is included that mostly lines up with the theory.	• An explanation is provided that clearly links up with the theory, but could have more detail.	• A comprehensive explanation is included that applies the theory.	
	0 – 4 Marks	**5 – 6 Marks**	**7 Marks**	**8 – 10 Marks**	

Marking Criteria	***Fail/Does not meet the required standard (0% to 49%)***	***Average/meets the required standard (50% to 64%)***	***Above average/is above the required standard (65% to 74%)***	**Excellent** ***(75% - 100%)***	**Feedback**
TASK 2 – CLASSIFICATION AND MODEL IMPROVEMENT					
Application: **An explanation is included of what the analysis is that will be conducted on the data set.** **[10 Marks]**	• No explanation included; • The explanation lacks details in terms of what the analysis will accomplish.	• An explanation is included that provides the bare basics of what the analysis will accomplish.	• An explanation is included that mostly explains what the analysis will be about, but could be more detailed.	• A well-motivated explanation is included that clearly details what the purpose of the analysis is.	
	0 – 4 Marks	**5 – 6 Marks**	**7 Marks**	**8 – 10 Marks**	
Application: **Python code is included that correctly performs the analysis.** **[40 Marks]**	• No Python code is included; • Python code does not run at all; • Python code only partially works.	• Python code mostly works correctly with one or two issues under normal usage scenarios.	• Python code works correctly under most usage scenarios.	• Python code works perfectly and will be able to handle any requests without issues.	
	0 – 19 Marks	**20 – 26 Marks**	**27 – 29 Marks**	**30 – 40 Marks**	

Marking Criteria	***Fail/Does not meet the required standard (0% to 49%)***	***Average/meets the required standard (50% to 64%)***	***Above average/is above the required standard (65% to 74%)***	**Excellent** ***(75% - 100%)***	**Feedback**
TASK 2 – CLASSIFICATION AND MODEL IMPROVEMENT					
Knowledge: **An explanation is included of the process used for the data analysis.** **[10 Marks]**	• No explanation is included; • The explanation is unclear; • The explanation is not relevant to the specific code.	• The explanation includes all the steps but lacks detail; • The explanation includes only some of the steps in detail.	• The explanation covers all the steps but lacks detail in some cases.	• The explanation is fully detailed, explaining every step of the process.	
	0 – 4 Marks	**5 – 6 Marks**	**7 Marks**	**8 – 10 Marks**	
Application: **An explanation of how the model was evaluated and improved** **[20 Marks]**	• No evidence is submitted of model evaluation and improvement; • The model was evaluated but not improved; • The model evaluation and improvement lacks detail.	• The model is evaluated and improved but the explanation lacks detail; • The model is evaluated and improved but the improvement was not noticeable.	• The model was evaluated and improved, most correctly applying the theory.	• A description of the model evaluation that applies the theory correctly and improves the performance of the model significantly.	
	0 – 9 Marks	**10 – 13 Marks**	**14 Marks**	**15 – 20 Marks**	

Marking Criteria	***Fail/Does not meet the required standard (0% to 49%)***	***Average/meets the required standard (50% to 64%)***	***Above average/is above the required standard (65% to 74%)***	**Excellent *(75% - 100%)***	**Feedback**
TASK 3 – TEXT PROCESSING AND PIPELINES					
Knowledge: **Explanation of what text processing is.** **[10 Marks]**	• No explanation included; • Explanation is unclear or incorrect.	• A correct but basic explanation is included.	• An explanation is included with additional details beyond the basics.	• A comprehensive explanation is included.	
	0 – 4 Marks	**5 – 6 Marks**	**7 Marks**	**8 – 10 Marks**	
Application: **Explanation of why the chosen data set is appropriate for analysis with text processing.** **[10 Marks]**	• No explanation is included; • The explanation does not link up with the theory at all.	• A basic explanation is included that links up with the theory; • An explanation is included that mostly lines up with the theory.	• An explanation is provided that clearly links up with the theory, but could have more detail.	• A comprehensive explanation is included that applies the theory.	
	0 – 4 Marks	**5 – 6 Marks**	**7 Marks**	**8 – 10 Marks**	

Marking Criteria	*Fail/Does not meet the required standard (0% to 49%)*	*Average/meets the required standard (50% to 64%)*	*Above average/is above the required standard (65% to 74%)*	**Excellent** *(75% - 100%)*	**Feedback**
TASK 3 – TEXT PROCESSING AND PIPELINES					
Application: **An explanation is included of what the analysis is that will be conducted on the data set.** **[10 Marks]**	• No explanation included; • The explanation lacks details in terms of what the analysis will accomplish.	• An explanation is included that provides the bare basics of what the analysis will accomplish.	• An explanation is included that mostly explains what the analysis will be about, but could be more detailed.	• A well-motivated explanation is included that clearly details what the purpose of the analysis is.	
	0 – 4 Marks	**5 – 6 Marks**	**7 Marks**	**8 – 10 Marks**	
Application: **Python code is included that correctly performs the analysis.** **[40 Marks]**	• No Python code is included; • Python code does not run at all; • Python code only partially works.	• Python code mostly works correctly with one or two issues under normal usage scenarios.	• Python code works correctly under most usage scenarios.	• Python code works perfectly and will be able to handle any requests without issues.	
	0 – 19 Marks	**20 – 26 Marks**	**27 – 29 Marks**	**30 – 40 Marks**	

Marking Criteria	***Fail/Does not meet the required standard (0% to 49%)***	***Average/meets the required standard (50% to 64%)***	***Above average/is above the required standard (65% to 74%)***	**Excellent** ***(75% - 100%)***	**Feedback**
TASK 3 – TEXT PROCESSING AND PIPELINES					
Application: **A pipeline is used in the Python code.** **[20 Marks]**	• No pipeline is used; • Some attempt is made at using a pipeline but it is not working correctly.	• Some use is made of a pipeline, but there are minor issues with the code.	• The pipeline is mostly working correctly.	• A pipeline is used to fully automate the whole analysis process.	
	0 – 9 Marks	**10 – 13 Marks**	**14 Marks**	**15 – 20 Marks**	
Knowledge: **An explanation is included of the process used for the data analysis.** **[10 Marks]**	• No explanation is included; • The explanation is unclear; • The explanation is not relevant to the specific code.	• The explanation includes all the steps but lacks detail; • The explanation includes only some of the steps in detail.	• The explanation covers all the steps but lacks detail in some cases.	• The explanation is fully detailed, explaining every step of the process.	
	0 – 4 Marks	**5 – 6 Marks**	**7 Marks**	**8 – 10 Marks**	

[TOTAL MARKS: 300]