



TAYLOR'S UNIVERSITY

Wisdom • Integrity • Excellence

**SCHOOL OF COMPUTING AND INFORMATION TECHNOLOGY
BACHELOR OF SOFTWARE ENGINEERING
BACHELOR OF COMPUTER SCIENCE**

GROUP ASSIGNMENT GUIDELINE

MODULE TITLE	:	DATA SCIENCE PRINCIPLES
MODULE CODE	:	ITS65704
WEIGHTAGE	:	30%
DUE DATE	:	WEEK 12

Note:

*** PLAGIARISM IS A SERIOUS OFFENCE AND PLAGIARIZED WORK WILL RESULT IN AN F GRADE.**

0 mark and a barring from sitting final examination will be implemented for those who does not submit any assignments.

ASSIGNMENT DETAILS:

A: Objective

The objective of this assignment is to use the Data Science processes learnt to analyse a data set and present the output in an interesting critical manner using python (Jupyter Notebook).

B: Instruction

1. This is a group assignment (3 members).
2. You will need to do this assignment on a computer with python (Jupyter Notebook) installed.

C: General Guidelines

The idea of the group project is to give you some experience trying to do a piece of original research in Data Science and writing up your results in a paper style format. What we expect to see is an idea/task that you describe clearly, relate to existing work, implement and test on a dataset. To do this you will need to write code, run it on some data, make some figures, read a few background papers, collect some references, and write a few pages describing your task, the algorithm(s) you used and the results you obtained.

The grade will depend on the ideas, how well you present them in the report, how clearly you position your work relative to existing literature, how illuminating your experiments are, and well-supported your conclusions are.

The idea is that this project report should be a manageable amount of work, but that if you want to turn your project into a paper, everything in the project report will need to be done anyway. If you feel that your project won't fit into this rubric, please talk to your instructor. There are many ways to make contributions to a field!

Your project should consist of an implementation of one or more machine learning algorithms, and their application to a dataset. Your project may be a comparison of several existing algorithms, or it may propose a new algorithm, which should then be compared to at least one other approach. Finding a Data Science project, dataset and applying Data Mining techniques to it can also be of interest. As a under graduate student, you are free to pick a project of your own design.

You are free to use any third-party ideas or algorithms that you wish as long as it is publicly available. You must properly provide references to any work that is not your own in the write-up. The project is not intended to be a stressful exercise; instead it is a chance for you to experiment, to think, to play and to hopefully have fun! Start with simple methods that work more or less out of the box and go from there.

D. Specific Requirements

Length: 4 to 8 pages, not including appendices. Don't be afraid to keep the text short and to the point, and to include large illustrative figures.

1. Abstract (5 points): a summary of the main idea of the project and its contributions.
 - a. Should be understandable to anyone in the course.
 - b. You don't need to say everything you did, just say what the main idea was and what were one or two takeaways.
2. Introduction (5 points): Introduce the topic of the project
3. Related work (10 points): A section describing related works and the bibliography.
 - a. If your project builds on previous work, clearly distinguish what they did from what your new contribution is.
 - b. Include a 1-2 sentence summary of other closely related papers. You might not know about all related papers (or have time to carefully read all related papers), and that's OK for this project. A rough guide is that you should be able to find 3-4 closely related papers, and another 3-4 papers that all those papers cite as foundational work. These foundational papers are often cited in the introduction.
4. Problem Statement (10 points): Include the answer for the following questions:
 - a. Describe how things should work
 - b. Explain the problem and state why it matters
 - c. Explain your problem's financial costs
 - d. Back up your claims
 - e. Propose a solution
 - f. Explain the benefits of your proposed solutions
 - g. Conclude by summarizing the problem and solution

5. Data Acquisition (10 points): Give the data acquisition process.
- Data is identified with use cases
 - Prospect of the required data is carried out
 - Consider the qualified data sources only
 - Consider the Ethical issues of the data
 - Split the data for training, testing and validation
 - Semantic analysis of the data sets is undertaken to understand the data features
6. Data Preparation (10 points): Give the procedures how the data was prepared for analysis
- Gather the data
 - Discover and assess the data
 - Clean and validate the data
 - Transform and enrich the data
 - Store the data
7. Exploratory Data Analysis (10 points): Give the steps taken to explore the data
- Description of the data
 - Handling missing data
 - Handling outliers
 - Understanding relationship and new insights through plots
 - Feature Extraction
 - Feature Selection
8. Model Development (10 points): Give the steps of in the model development
- Model selection
 - Model fitting
 - Model validation

9. Visualization and Communication (10 points); Use the appropriate presentation techniques and tools to visualize the results
 - a. Communicate the results
 - b. Determine the best method/graph
10. Deploy and maintain (10 points): Deploy and maintain the model
 - a. Testing
 - b. Validation
 - c. Improvement
11. Limitations (5 points): A section describing the limitations of your approach.
 - a. Describe some settings in which we'd expect your approach to perform poorly, or where all existing models fail.
 - b. Try to guess or explain why these limitations are the way they are.
 - c. Give some examples of possible extensions, ways to address these limitations, or open problems.
12. Conclusions (5 points): A section describing your conclusions and ideas for future work.
 - a. State the results achieved in relation to the problem described in the introduction.
 - b. Repeat the main takeaways from your paper.

E: Format of the Written Report

1. You are required to use the python (Jupyter Notebook).
2. You are required to compile your report in the format below:
 - a. Students are required to submit the softcopy (ipynb) of the report.
 - b. Students are required to submit softcopy (via ***submission link***) of the report.
 - c. Due Date: **Week 12 Practical Class**

No late submission is accepted (except for strong valid reason and prior approval granted by the lecturer).

D: Assessment Marking Criteria

Content

1. Abstract	: 5
2. Introduction	: 5
3. Related Work	: 10
4. Problem Statement	: 10
5. Data Acquisition	: 10
6. Data Preparation	: 10
7. Exploratory Data Analysis	: 10
8. Model Development	: 10
9. Visualization and Communication	: 10
10. Deploy and Maintain	: 10
11. Limitation	: 5
12. Conclusion	: 5

Total : 100

Marking Rubric for Written Assignment (Group):

No	Criteria						
1.	Abstract (5 Points)	5 marks A clear and concise abstract that gives the reader a clear idea of what the project is about and why it is interesting. The following components need to be included <ul style="list-style-type: none"> i. Purpose and motivation of this research ii. Problem you are addressing iii. Methods and materials iv. Results v. Conclusion 	4 marks A clear abstract that gives the reader a clear idea of what the project is about. Four of the following components are included <ul style="list-style-type: none"> i. Purpose and motivation of this research ii. Problem you are addressing iii. Methods and materials iv. Results v. Conclusion 	3 marks The abstract is difficult to read and/or is very vague and/or doesn't sell the project as well as it might have. Three of the following components are included <ul style="list-style-type: none"> i. Purpose and motivation of this research ii. Problem you are addressing iii. Methods and materials iv. Results v. Conclusion 	2 marks Unable to read the abstract and/or is very vague and/or doesn't sell the project as well as it might have. Only two of the following components are included <ul style="list-style-type: none"> i. Purpose and motivation of this research ii. Problem you are addressing iii. Methods and materials iv. Results v. Conclusion 	1 mark Unable to read the abstract. Only one of the following components is included <ul style="list-style-type: none"> i. Purpose and motivation of this research ii. Problem you are addressing iii. Methods and materials iv. Results v. Conclusion 	
2.	Introduction (5 Points)	5 marks A readable write-up that explains what the problem is and why it is of interest. The following components need to be included <ul style="list-style-type: none"> i. Problem ii. Negative impact of the problem iii. Parties affected iv. Benefit of solving the problem 	4 marks A readable write-up that explains what the problem is. Three of the following components are included. <ul style="list-style-type: none"> i. Problem ii. Negative impact of the problem iii. Parties affected iv. Benefit of solving the problem 	3 marks The write-up is difficult to read, somewhat vague, or doesn't make a really good case for why the problem is of interest. Two of the following components are included. <ul style="list-style-type: none"> i. Problem ii. Negative impact of the problem iii. Parties affected iv. Benefit of solving the problem 	2 marks Unable to read the write-up and/or is very vague. Only one of the following components are included. <ul style="list-style-type: none"> i. Problem ii. Negative impact of the problem iii. Parties affected iv. Benefit of solving the problem 	1 mark Unable to read the write-up. None of the following components are included. <ul style="list-style-type: none"> i. Problem ii. Negative impact of the problem iii. Parties affected iv. Benefit of solving the problem 	
3.	Related work (10 points)	9-10 marks An outstanding overview, with an insightful analysis of prior work and a clear	7-8 marks A comprehensive overview of prior work that gives the reader a	5-6 marks A fairly good overview of prior work, and some connection is made to	3-4 marks An overview of several papers related to the proposed method, and	1-2 marks Bad attempt at describing prior work. None of the	

		<p>connection between prior work and the proposed method. The following components are given.</p> <ol style="list-style-type: none"> Introduction of the topic Taxonomy Mapping Paragraphs for each branch of the taxonomy tree Conclusion Critical Review 	<p>clear idea of what's out there and how the proposed method is different. Four of the following components are given.</p> <ol style="list-style-type: none"> Introduction of the topic Taxonomy Mapping Paragraphs for each branch of the taxonomy tree Conclusion Critical Review 	<p>the proposed method. Three of the following components are given.</p> <ol style="list-style-type: none"> Introduction of the topic Taxonomy Mapping Paragraphs for each branch of the taxonomy tree Conclusion Critical Review 	<p>some attempt is made to connect the prior work to the current method. Two of the following components are given.</p> <ol style="list-style-type: none"> Introduction of the topic Taxonomy Mapping Paragraphs for each branch of the taxonomy tree Conclusion Critical Review 	<p>following components are given.</p> <ol style="list-style-type: none"> Introduction of the topic Taxonomy Mapping Paragraphs for each branch of the taxonomy tree Conclusion Critical Review 	
4.	Problem Statement (10 Points)	<p>9-10 marks A thought-out, clear, and original illustration that makes the idea immediately clear. The following components need to be included</p> <ol style="list-style-type: none"> Describe how things should work Explain the problem and state why it matters Explain your problem's financial costs Back up your claims 	<p>7-8 marks An illustration that does the job, but is not particularly clear or original. The following components need to be included</p> <ol style="list-style-type: none"> Describe how things should work Explain the problem and state why it matters Explain your problem's financial costs Back up your claims 	<p>5-6 marks A somewhat incomplete description where details can be reconstructed with some effort. The following components need to be included</p> <ol style="list-style-type: none"> Describe how things should work Explain the problem and state why it matters Explain your problem's financial costs Back up your claims 	<p>3-4 marks An illustration that is significantly lacking in some respect. The following components need to be included</p> <ol style="list-style-type: none"> Describe how things should work Explain the problem and state why it matters Explain your problem's financial costs Back up your claims Propose a solution 	<p>1-2 marks An illustration that is poorly written. The following components need to be included</p> <ol style="list-style-type: none"> Describe how things should work Explain the problem and state why it matters Explain your problem's financial costs Back up your claims Propose a solution 	

		v. Propose a solution vi. Explain the benefits of your proposed solutions vii. Conclude by summarizing the problem and solution	v. Propose a solution vi. Explain the benefits of your proposed solutions vii. Conclude by summarizing the problem and solution	v. Propose a solution vi. Explain the benefits of your proposed solutions vii. Conclude by summarizing the problem and solution	vi. Explain the benefits of your proposed solutions vii. Conclude by summarizing the problem and solution	vi. Explain the benefits of your proposed solutions vii. Conclude by summarizing the problem and solution	
5.	Data Acquisition (10 points)	9-10 marks An insightful and correct steps of data acquisition. The following components are given. i. Data is identified with use cases ii. Prospect of the required data is carried out iii. Consider the qualified data sources only iv. Consider the Ethical issues of the data v. Split the data for training,	7-8 marks A correct steps of data acquisition that could be more complete and is not very insightful. One of the following components is missing. i. Data is identified with use cases ii. Prospect of the required data is carried out iii. Consider the qualified data sources only iv. Consider the Ethical issues of the data v. Split the data for training,	5-6 marks An incomplete or somewhat incorrect steps of data acquisition. Two of the following components are missing. i. Data is identified with use cases ii. Prospect of the required data is carried out iii. Consider the qualified data sources only iv. Consider the Ethical issues of the data	3-4 marks An incorrect steps of data acquisition. One of the following components are given. i. Data is identified with use cases ii. Prospect of the required data is carried out iii. Consider the qualified data sources only iv. Consider the Ethical issues of the data v. Split the data for training,	1-2 marks No steps of data acquisition. None of the following components are given. i. Data is identified with use cases ii. Prospect of the required data is carried out iii. Consider the qualified data sources only iv. Consider the Ethical issues of the data v. Split the data for training,	

		<p>testing and validation</p> <p>vi. Semantic analysis of the data sets is undertaken to understand the data features</p>	<p>testing and validation</p> <p>vi. Semantic analysis of the data sets is undertaken to understand the data features</p>	<p>v. Split the data for training, testing and validation</p> <p>vi. Semantic analysis of the data sets is undertaken to understand the data features</p>	<p>testing and validation</p> <p>vi. Semantic analysis of the data sets is undertaken to understand the data features</p>	<p>testing and validation</p> <p>vi. Semantic analysis of the data sets is undertaken to understand the data features</p>	
6.	Data Preparation (10 points)	<p>9-10 marks An insightful and correct steps of data preparation. The following components are given.</p> <p>i. Gather the data</p> <p>ii. Discover and assess the data</p> <p>iii. Clean and validate the data</p> <p>iv. Transform and enrich the data</p> <p>v. Store the data</p>	<p>7-8 marks A correct steps of data preparation that could be more complete and is not very insightful. One of the following components is missing.</p> <p>i. Gather the data</p> <p>ii. Discover and assess the data</p> <p>iii. Clean and validate the data</p> <p>iv. Transform and enrich the data</p> <p>v. Store the data</p>	<p>5-6 marks An incomplete or somewhat incorrect steps of data preparation. Two of the following components are missing.</p> <p>i. Gather the data</p> <p>ii. Discover and assess the data</p> <p>iii. Clean and validate the data</p> <p>iv. Transform and enrich the data</p> <p>v. Store the data</p>	<p>3-4 marks An incorrect steps of data preparation. One of the following components are given.</p> <p>i. Gather the data</p> <p>ii. Discover and assess the data</p> <p>iii. Clean and validate the data</p> <p>iv. Transform and enrich the data</p> <p>v. Store the data</p>	<p>1-2 marks No steps of data preparation. None of the following components are given.</p> <p>i. Gather the data</p> <p>ii. Discover and assess the data</p> <p>iii. Clean and validate the data</p> <p>iv. Transform and enrich the data</p> <p>v. Store the data</p>	

7.	Exploratory Data Analysis (EDA) (10 points)	9-10 marks An insightful and correct set of steps of Exploratory Data Analysis. The following components are given. <ul style="list-style-type: none"> i. Description of the data ii. Handling missing data iii. Handling outliers iv. Understanding relationship and new insights through plots v. Feature Extraction vi. Feature Selection 	7-8 marks A set of correct steps of Exploratory Data Analysis that could be more complete and is not very insightful. One of the following components is missing. <ul style="list-style-type: none"> i. Description of the data ii. Handling missing data iii. Handling outliers iv. Understanding relationship and new insights through plots v. Feature Extraction vi. Feature Selection 	5-6 marks An incomplete or somewhat incorrect set of steps of Exploratory Data Analysis. Two of the following components are missing. <ul style="list-style-type: none"> i. Description of the data ii. Handling missing data iii. Handling outliers iv. Understanding relationship and new insights through plots v. Feature Extraction vi. Feature Selection 	3-4 marks An incorrect set of steps of Exploratory Data Analysis. One of the following components are given. <ul style="list-style-type: none"> i. Description of the data ii. Handling missing data iii. Handling outliers iv. Understanding relationship and new insights through plots v. Feature Extraction vi. Feature Selection 	1-2 marks No steps of Exploratory Data Analysis. None of the following components are given. <ul style="list-style-type: none"> i. Description of the data ii. Handling missing data iii. Handling outliers iv. Understanding relationship and new insights through plots v. Feature Extraction vi. Feature Selection 	
8.	Model Development (10 points)	9-10 marks An insightful and correct set of steps of model development. The following components are given. <ul style="list-style-type: none"> i. Algorithm selection ii. Model selection 	-8 marks A set of correct steps of model development that could be more complete and is not very insightful. One of the following components is missing. <ul style="list-style-type: none"> i. Algorithm selection ii. Model selection 	An incomplete or somewhat incorrect set of steps of model development. Two of the following components are missing. <ul style="list-style-type: none"> i. Algorithm selection ii. Model selection 	3-4 marks An incorrect set of steps of model development. One of the following components are given. <ul style="list-style-type: none"> i. Algorithm selection ii. Model selection iii. Model fitting 	1-2 marks No steps of model development. None of the following components are given. <ul style="list-style-type: none"> i. Algorithm selection ii. Model selection iii. Model fitting 	

		iii. Model fitting iv. Model validation	iii. Model fitting iv. Model validation	iii. Model fitting iv. Model validation	iv. Model validation	iv. Model validation	
9.	Visualization and Communication (10 points)	9-10 marks An insightful and correct Visualization and Communication analysis. The following components are given. <ul style="list-style-type: none"> i. Communicate the results ii. Determine the best method/graph iii. Frequency or probability distribution table 	7-8 marks A correct Visualization and Communication analysis that could be more complete and is not very insightful. One of the following components is missing. <ul style="list-style-type: none"> i. Communicate the results ii. Determine the best method/graph iii. Frequency or probability distribution table 	5-6 marks An incomplete or somewhat incorrect Visualization and Communication analysis. Two of the following components are missing <ul style="list-style-type: none"> i. Communicate the results ii. Determine the best method/graph iii. Frequency or probability distribution table 	-4 marks An incorrect Visualization and Communication analysis. One of the following components are given. <ul style="list-style-type: none"> i. Communicate the results ii. Determine the best method/graph iii. Frequency or probability distribution table 	1-2 marks No Visualization and Communication analysis. None of the following components are given. <ul style="list-style-type: none"> i. Communicate the results ii. Determine the best method/graph iii. Frequency or probability distribution table 	
10.	Deploy and Maintain (10 points)	9-10 marks An insightful and correct Deployment and Maintenance analysis. The following components are given. <ul style="list-style-type: none"> i. Testing ii. Validation iii. Improvement iv. Compare with other models 	7-8 marks A correct Deployment and Maintenance analysis that could be more complete and is not very insightful. One of the following components is missing. <ul style="list-style-type: none"> i. Testing ii. Validation iii. Improvement 	5-6 marks An incomplete or somewhat incorrect Deployment and Maintenance analysis. Two of the following components are missing <ul style="list-style-type: none"> i. Testing ii. Validation iii. Improvement 	-4 marks An incorrect Deployment and Maintenance analysis. One of the following components are given. <ul style="list-style-type: none"> i. Testing ii. Validation iii. Improvement iv. Compare with other models 	1-2 marks No Deployment and Maintenance analysis. None of the following components are given. <ul style="list-style-type: none"> i. Testing ii. Validation iii. Improvement iv. Compare with other models 	

		v. Compare with other studies	iv. Compare with other models v. Compare with other studies	iv. Compare with other models v. Compare with other studies	v. Compare with other studies	v. Compare with other studies	
11.	Limitations (5 points)	5 marks An insightful and correct analysis. The following components are given. <ul style="list-style-type: none"> i. Identify the limitation or limitations ii. Explain these limitations in detail iii. Propose a future direction for future studies 	4 marks A correct analysis that could be more complete and is not very insightful. One of the following components is missing. <ul style="list-style-type: none"> i. Identify the limitation or limitations ii. Explain these limitations in detail iii. Propose a future direction for future studies 	3 marks An incomplete or somewhat incorrect analysis. Two of the following components are missing. <ul style="list-style-type: none"> i. Identify the limitation or limitations ii. Explain these limitations in detail iii. Propose a future direction for future studies 	2 marks An incorrect analysis. One of the following components are given. <ul style="list-style-type: none"> i. Identify the limitation or limitations ii. Explain these limitations in detail iii. Propose a future direction for future studies 	1 marks No analysis. None of the following components are given. <ul style="list-style-type: none"> i. Identify the limitation or limitations ii. Explain these limitations in detail iii. Propose a future direction for future studies 	
12.	Conclusions (5 points)	5 marks A clear and insightful summary of the paper, perhaps with interesting ideas for future work. The following components are given. <ul style="list-style-type: none"> i. Restate your research topic ii. Restate the objective iii. Summarize the main topics iv. Significance of results v. Conclude the thoughts 	4 marks A summary of the experiments is given, but the conclusion is a mere summary. The ideas for future work are not interesting. One of the following components is missing. <ul style="list-style-type: none"> i. Restate your research topic ii. Restate the objective iii. Summarize the main topics iv. Significance of results 	3 marks A flawed conclusion. Two of the following components are missing. <ul style="list-style-type: none"> i. Restate your research topic ii. Restate the objective iii. Summarize the main topics iv. Significance of results v. Conclude the thoughts 	2 marks An incorrect conclusion. Three of the following components are missing. <ul style="list-style-type: none"> i. Restate your research topic ii. Restate the objective iii. Summarize the main topics iv. Significance of results v. Conclude the thoughts 	1 marks No conclusion. One of the following components is given. <ul style="list-style-type: none"> i. Restate your research topic ii. Restate the objective iii. Summarize the main topics iv. Significance of results v. Conclude the thoughts 	

			v. Conclude the thoughts				
--	--	--	--------------------------	--	--	--	--

END