

## University of Westminster School of Computer Science

7BUI5008W Data Mining & Machine Learning	
Module leader	Mahmoud Aldraimli
Unit	Coursework
Weighting:	50%
Qualifying mark	40%
Description	Students are expected to critically engage in effectively applying and evaluating novel data mining and machine learning techniques for a specific problem domain and definitely reflect on the knowledge of how different data mining and machine learning algorithms perform in terms of biases for a given problem domain. Students are expected to methodically analyse the output of the data mining tasks and machine learning algorithms by drawing technically appropriate and sound conclusions resulting from the application of data mining and machine learning algorithms to the given problem.
Learning Outcomes Covered in this Assignment:	<p>This assignment contributes towards the following Learning Outcomes (LOs):</p> <ul style="list-style-type: none"> <li>• LO2 fully implement data mining/machine learning projects, focused on problem analysis, data pre-processing, data post-processing by choosing and implementing appropriate algorithms;</li> <li>• LO4 fully implement encode and test data mining and machine learning algorithms using the programming language (such as Python) and standard packages and toolkits.</li> <li>• LO6 perform a critical evaluation of performance metrics for data mining and machine learning algorithms for a given domain/application.</li> </ul>
Handed Out:	16 <sup>th</sup> Oct 2024
Due Date	26 <sup>th</sup> November 2024 Submission by 13:00 hours
Expected deliverables	Submit your Word document report with the results/analysis on Blackboard. And submit a Python Notebook file containing the required implemented codes in Python notebook format (.ipynb).
Method of Submission:	Electronic submission on BB via a provided link close to the submission time.
Type of Feedback and Due Date:	Feedback will be provided on BB, the week starting 20 <sup>th</sup> December 2024
BCS CRITERIA MEETING IN THIS ASSIGNMENT	<ul style="list-style-type: none"> <li>• 7.1.6 Use appropriate processes</li> <li>• 7.1.7 Investigate and define a problem</li> <li>• 7.1.8 Apply principles of supporting disciplines</li> <li>• 8.1.1 Systematic understanding of knowledge of the domain with depth in particular areas</li> <li>• 8.1.2 Comprehensive understanding of essential principles and practices</li> <li>• 8.2.2 Tackling a significant technical problem</li> <li>• 10.1.2 Comprehensive understanding of the scientific techniques</li> </ul>

Refer to section 4 of the “How you study” guide for undergraduate students for a clarification of how you are assessed, penalties and late submissions, what constitutes plagiarism etc.

### **Penalty for Late Submission**

If you submit your coursework late but within 24 hours or one working day of the specified deadline, 10 marks will be deducted from the final mark as a penalty for late submission, except for work which obtains a mark in the range of 50 – 59%, in which case the mark will be capped at the pass mark (50%). If you submit your coursework more than 24 hours or more than one working day after the specified deadline, you will be given a mark of zero for the work in question unless a claim of Mitigating Circumstances has been submitted and accepted as valid.

It is recognised that, on occasion, illness or a personal crisis can mean that you fail to submit a piece of work on time. In such cases, you must inform the Campus Office in writing on a mitigating circumstances form, giving the reason for your late or non-submission. You must provide relevant documentary evidence with the form. This information will be reported to the relevant Assessment Board, which will decide whether the mark of zero shall stand. For more detailed information regarding University Assessment Regulations, please refer to the following website: <http://www.westminster.ac.uk/study/current-students/resources/academicregulations>

## **Coursework Description**

### **The Real-world Problem Description**

#### **A) The Domain**

The deployment of machine learning modelling in this coursework aims to tackle a real-world tool by developing effective early screening machine learning models for breast cancer mortality and survival prediction to help doctors enhance their treatment planning and management.

*Cancer* is a disease in which cells in the body grow out of control. Breast cancer is a disease in which abnormal breast cells grow out of control and form tumours. If left unchecked, the tumours can spread throughout the body and become fatal. Breast cancer cells begin inside the milk ducts and/or the milk-producing lobules of the breast.

In females in the UK, breast cancer is the 2nd most common cause of cancer death, with around 11,400 deaths every year (2017-2019). In males in the UK, breast cancer is not among the 20 most common causes of cancer death, with around 85 deaths every year (2017-2019).

#### **Stages and grades of breast cancer**

The tests and scans the patient have to diagnose breast cancer give information about:

- the size of the cancer and whether it has spread (the stage)
- how abnormal the cells look under the microscope (the grade)

Knowing the stage and grade helps doctor plan the patient’s treatment. The stage of a cancer tells the patient how big it is and whether it has spread. It helps the doctor decide which treatment the patient need.

There are different systems used in the UK to stage breast cancer. The most common one is the TNM system. TNM stands for Tumour, Node and Metastasis. the patient might also be told about the number staging system. There are 4 main stages in this system, from 1 to 4.

The tests the patient has also give information about the type of breast cancer they have.

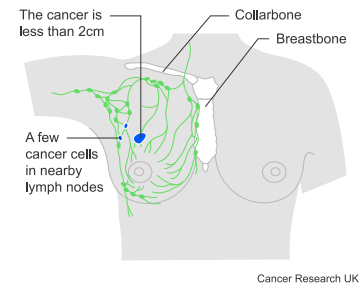
The information below is an overview of the TNM staging for all types of cancer.

- **T** describes the size of the tumour (cancer)
- **N** describes whether there are any cancer cells in the nearby lymph nodes
- **M** describes whether the cancer has spread to parts of the body further away from where the cancer started

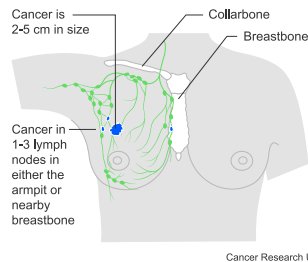
**Stage 1** breast cancer means that the cancer is small and only in the breast tissue or it might be found in lymph nodes close to the breast (see Figure 1). It is an early-stage breast cancer.

The stage of cancer tells the patient how big it is and how far it has spread. It helps the doctor decide the best treatment for the patient. There are different systems used in the UK to stage breast cancer. Stage 1 is part of the number staging system. Doctors may also use the TNM staging system.

Staging for breast cancer is very complex. Many different factors are considered before doctors can confirm the patient's final stage.



**Fig.1** Illustration of stage 1 breast cancer

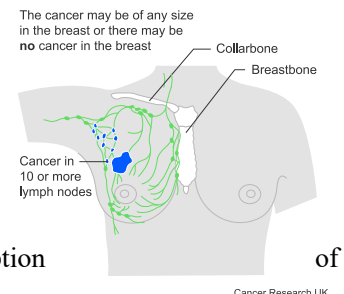


**Fig.2** Illustration of stage 2 breast cancer

**Stage 2** breast cancer means that the cancer is either in the breast or in the nearby lymph nodes or both. It is an early-stage breast cancer.

Stage 2 is part of the number staging system. Doctors may also use the TNM staging system.

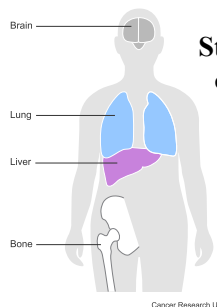
Stage 2 can be divided into 2A and 2B. Opposite is a simplified description of stage 2A and 2B breast cancer (see Figure 2).



**Fig.3** Illustration of stage 3 breast cancer

**Stage 3** means that the cancer has spread from the breast to the lymph nodes close to the breast, to the skin of the breast or to the chest wall. It is also called locally advanced breast cancer. Stage 3 is part of the number staging system. Doctors may also use the TNM staging system.

Stage 3 can be divided into 3A, 3B and 3C. opposite is a simplified description of stage 3A, 3B and 3C breast cancer (see Figure 3).



**Fig.4** Illustration of stage 4 breast cancer

**Stage 4** breast cancer has spread to another part of the body (see Figure 4). It is also called advanced cancer or secondary breast cancer. The aim of treatment is to control the cancer and any symptoms. Treatment depends on a number of factors.

In stage 4 breast cancer:

- the cancer can be any size
- the lymph nodes may or may not contain cancer cells
- the cancer has spread (metastasised) to other parts of the body such as the bones, lungs, liver or brain.

## Hormonal Treatment

Hormone therapy is a common treatment for secondary breast cancer. It can often shrink and control the cancer wherever it is in the body. It works well if the cancer cells have particular proteins called hormone receptors, estrogen receptor, progesterone receptor.

If one hormone therapy stops working so well, the doctor might suggest you try a different one.

## Other Treatment

Doctor will take many different factors into account when deciding which treatment is best for the patient. These include:

- the type of cells the cancer started in
- which part of your body the cancer has spread to
- the treatment you have already had
- your general health
- whether the patient have had the menopause.

- whether the cancer is growing slowly or more quickly
- whether the cancer cells have receptors for particular cancer drugs

If your cancer doesn't have hormone receptors or has spread to the liver or lungs, the doctor might suggest Chemotherapy. Radiotherapy might be recommended if the cancer has spread to the bones or the skin near the breast. Targeted and immunotherapy drugs might be recommended for secondary breast cancer.

### C) The Domain Problem

The importance of predicting mortality, short- and long-term survival of patients with cancer may improve their care. Prior predictive models either use data with limited availability or predict the outcome of only 1 type of cancer. In this case breast cancer.

### D) Your Role as A Data Scientist

You are hired as a data scientist to work alongside a team of doctors to

- 1- Build predictive machine-learning models for breast cancer mortality status.
- 2- Build predictive machine-learning models to estimate patient's survival period.

The team of doctors provided you with historical records of breast cancer patients and had their mortality status. Also, obtained the number of months they survived.

The doctors rely on your work to answer the following **two research question** on the dataset; the key objective is to create a new, predictive tool powered by a machine learning model to assist doctors in enhancing their treatment planning and cancer care. **The Research Questions are:**

- a) Does machine learning have the potential to assist doctors to predict those who would survive breast cancer or not?
- b) For patients who would not survive cancer, can machine learning offer a reliable estimate of their survival period?

### E) Your Dataset

This dataset of breast cancer patients was obtained from the 2017 November update of the SEER Program of the NCI, which provides information on population-based cancer statistics. The dataset contains the following attributes:

**Table.1** Data Dictionary

Attribute	Description
Patient ID	Unique identification for each patient
Month of Birth	A patient's month of birth
Age	A patient's month of birth in years
Sex	A patient's genomic sex
Race	Patient's ethnicity group
Marital Status	Married, Single, Divorced, Separated, Widowed
Occupation Code	Patient's job role
Adopted Status	0: Not adopted, 1: Adopted as a child
T Stage	The T stage in breast cancer refers to the size of the tumour from T1, T2, T3 and T4
N Stage	Used to indicate if the breast cancer has spread to surrounding lymph nodes (N), with a higher number representing a greater number of lymph nodes impacted, from N1, N2 and N3.
6th Stage	Breast Imaging Reporting and Data System or BI-RADS
Differentiate	How the cancer cells look and are growing compared with normal cells.
Grade	Breast Cancer Grades (Nottingham Grade)
A Stage	Breast cancer is staged based on how far it has spread.

	<b>Regional:</b> The cancer has spread to nearby lymph nodes or tissues. <b>Distant:</b> The cancer has spread to distant parts of the body, such as the lungs, liver, or bones
<b>Tumour Size</b>	Tumor size measured in millimeters
<b>Estrogen Status</b>	Cancer cells have estrogen hormone receptors or not.
<b>Progesterone Status</b>	Cancer cells have progesterone hormone receptors or not.
<b>Regional Node Examined</b>	Count of examined regional lymph nodes for cancer spread
<b>Regional Node Positive</b>	Count of cancer positive regional lymph nodes to contain metastases
<b>Survival Months</b>	Survival months based on date of last contact.
<b>Mortality Status</b>	Any patient that dies after the follow-up cut-off date is recoded to alive as of the cut-off date. If date of last contact > study cutoff date, vital status recode = alive.

**Note:** for general knowledge only and not for the purpose of this coursework, further information about the dataset can be found in <https://ieee-dataport.org/open-access/seer-breast-cancer-data>  
Survival calculations can be found in <https://seer.cancer.gov/survivaltime/>  
And <https://seer.cancer.gov/survivaltime/SurvivalTimeCalculation.pdf>

## Your Coursework Tasks & Framework

As a data scientist, you are a logician, a mathematician, a technician, and an analyst, and you need doctors to understand your analyses. Doctors are usually busy individuals, and they don't have all the time in the world. One essential skill that you must adhere to is to **be concise and straight to the point**. Focus on the answers needed for each task, and **provide just enough words for the answer only**. There is no need to provide lengthy descriptions of algorithms and methods unless you are asked to do.

Also, they are only interested in assessing your interpretation of the modelling results, so **you MUST NOT paste any Python code** in this report **unless specifically asked to so**. You will receive a separate link to submit your code as a Python notebook file (mandatory). **ipynb extension**. Your data mining tasks will be aligned with the popular CRISP-DM methodology phases but without the deployment phase (see Figure 5).

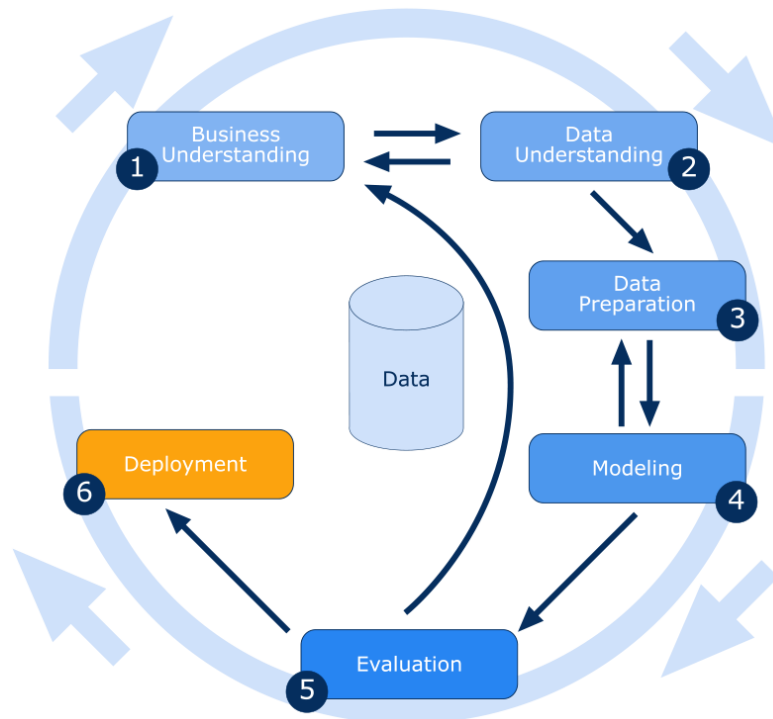




Fig.2 CRISP-DM Phases

 **Important Note:** You must answer each task chronologically and use the questions as headers 

## PART (A) Breast Cancer Mortality Prediction [65 MARKS]

### For Research Question A

Does machine learning have the potential to assist doctors to predict those who would survive breast cancer or not

**Task (1) – Domain Understanding: Classification**

**[Total 6 Marks]**

The doctors decided that classification modelling is required. Indicate in the table below for each of the listed variables in your data which ones you should RETAIN and can be included in the classification modelling of Breast Cancer Mortality (Alive vs. Dead) and the variables you should DROP (REMOVE). Justify your decision logically and/or by research (include in-text citation)

Variable Name	Retain or Drop	Brief justification for retention or dropping
Patient ID		
Month of Birth		
Age		
Sex		
Race		
Marital Status		
Occupation Code		
Adopted Status		
T Stage		
N Stage		
6th Stage		
Differentiate		
Grade		
A Stage		
Tumour Size		
Estrogen Status		
Progesterone Status		
Regional Node Examined		
Regional Node Positive		
Survival Months		
Mortality Status		

**Task (2) – Data Understanding: Producing Your Experimental Designing****[Total 3 Marks]**

From your Python notebook, for your RETAINED input variables and your class “target” variable, produce a basic statistical description and variable scale type. Plot the distribution of your target variable. (Paste screenshots of code **OUTPUTS ONLY** for evidence).

**Task (3) – Data Preparation: Cleaning and Transforming your data****[Total 16 Marks]**

a) Investigate any issues in your retained dataset and the possible variables. Based on the issues you find in your data, suggest a suitable possible method to mitigate each of these issues and provide your justification for using each method. Use the table below to organise your findings, add more rows if needed:

**[8 Marks]**

Variable Name	Issue description	Proposed mitigation	Justification for used mitigation
⋮	⋮	⋮	⋮

b) With the aid of Python packages and a notebook, **implement your suggested mitigations** of issues in Task (3.a), and **show evidence** of implementing your suggested solutions to the problems you identified for your dataset in (Task 3. a). (Use screenshots of code **OUTPUTS ONLY**). **Indicate and annotate in your screenshots** which issue was resolved from each screenshot provided. Show screenshots of code outputs before and after implementing your solution.

**[8 Marks]****Task (4) – Modelling: Create Predictive Classification Models****[Total 10 Marks]**

a) From the classification algorithms which you learned in the module, four different algorithms were selected: **Logistic Regression (LR)**, **K Nearest Neighbour (KNN)** and **Naïve Bayes (NB)**. These algorithms are a mix of parametric and non-parametric algorithms. List down the type of each algorithm (parametric vs non-parametric), name any learnable parameters, and list any possible hyperparameters for each algorithm which you may want to consider tuning. Note the Python package and module for importing each algorithm. Again, organise your answer in a table as before. See below:

**[6 Marks]**

Algorithm Name	Algorithm Type	Learnable Parameters	Some Possible Hyperparameters	Imported Python package to use the algorithm
NB				
LR				
KNN (N=?)				

b) With the aid of the Python packages, use the training–test split approach with your retained applicable categorical input features only and the class output feature to build your predictive classification models.

**[4 Marks]**

i. Screenshot the list of **all feature names used for building your classification models** and the corresponding **data shape** function output.

ii. In less than 100 words, research and justify your **choice of the training–test split ratio** and provide an in-text citation.

iii. In less than 100 words, discuss the overall purpose of using a training-test approach in contrast to the use of validation sets in K-fold cross-validation and describe the case/s when to apply each of those approach is used.

iv. Provide as evidence the code line from your source code that ensures that all models were tested on the same test dataset, also ensure that the labels ratio of Mortality Status “Alive” to Mortality Status “Dead” is the same in the training and test sets.

**Task (5) – Evaluation: How good are your models**

**[Total 30 Marks]**

Your healthcare professionals provided the following success criteria to guide you when evaluating your models.

*“When evaluating your model's performance, which addresses your first research question (a). The model is expected to misclassify subjects. Thus, the model should aim to predict the “Dead” mortality status of subjects for as many as possible to increase the urgency in treatment planning. However, the model should demonstrate that its high “Death” mortality prediction rate is mainly due to a larger portion of correctly detected (predicted) subjects who belong to the Mortality Status “Dead” class.”*

a) With the aid of Python packages, paste the test confusion matrix for each trained model as screenshots from the output of your Python code. [3 marks]

b) Five different classification evaluation metrics are noted. Paste each model's test performance results. State which evaluation metric/metrics to “USE or “NOT USE” to closely interpret the above success criteria. For justification, explain how closely your choice of “USE” or “DO NOT USE” for a metric interprets the given success criteria. With the aid of Python packages, document the TEST SCORES for each built model. [15 marks]

Metrics	USE or DO NOT USE	Justification in relation to the success criteria	Model Name	Test Score
Accuracy			NB	
			LR	
			KNN (K=?)	
Recall			NB	
			LR	
			KNN (K=?)	
Precision			NB	
			LR	
			KNN (K=?)	
F-Score			NB	
			LR	
			KNN (K=?)	
AUC-ROC			NB	
			LR	
			KNN (K=?)	



- c) Suggest a single best classification model based on the 'USED' performance metrics scores you identified in (Task 5. b). Briefly describe how well your best model satisfies the needs of your healthcare professionals. [2 marks]
- d) Investigated with evidence to establish whether your selected best model is good fit, underfit or overfit. [3 marks]
- e) To enhance your selected best model/s performance, tune some of its possible hyperparameters, which you indicated in (Task 4. a) for that specific algorithm. With the aid of Python packages, Re-train the algorithm again with GridSearchCV [5 marks]
- i. Indicate the number of cross-validation K folds used.
- ii. For the newly tuned model, document the estimated best hyperparameters,
- iii. Present the test confusion matrix for the best models before and after tuning.
- iv. Calculate and document the new score/s of the "USED" performance metric/s of your choice to interpret the success criteria identified in (Task 5.b) before and after tuning.
- v. Use your observations to **comment on whether the tuning of hyperparameters of your best model improved its positive predictive ability** in line with the success criteria.
- f) Based on your best model, draft an answer for the research question, criticise your best-performing model, and state any limitations you may have identified. Research and try to explain why your selected algorithm overtook all other models in no more than 100 words. State any ethical issues your model may raise if used to screen for breast cancer mortality. [2 Marks]

---

## PART (B) Breast Cancer Survival Rate Prediction [35 Marks]

### For Research Question B

For patients who would not survive breast cancer, can machine learning offer a reliable estimate of their survival period?

---

#### Task (1) – Domain Understanding: Regression

[Total 2 Marks]

The doctors decided that regression modelling is required. Using python functions, show the dimensions of your data subset that you will RETAIN for this regression modelling problem. Using python functions, list the names of the features that you intend to use for modelling from Table.1.

#### Task (2) – Data Understanding: Producing Your Experimental Designing [Total 5 Marks]

From your Python notebook, Plot the distribution for your RETAINED input variables and your "target" variable, (Paste screenshots of code **OUTPUTS ONLY** "the plots" for evidence).

#### Task (3) – Data Preprocessing: Transforming your data

[Total 5 Marks]

- a) By looking at the dataset establish whether there is a need for scaling your dataset attributes. Explain with evidence from your python code output the reasoning behind your recommendation. [2 marks]
- b) In general, when applying scaling to any dataset for regression modelling, would scale the input features only, the target feature only, or all features? In less than 150 words, briefly justify your answer and include in-text citation where appropriate. [3 marks]

#### Task (4) – Modelling: Build Predictive Regression Models

[Total 7 Marks]

- a) From the regression algorithms which you learned in the module, doctors decided on the use of a **Decision Tree Regression (DT)** algorithm. In less than 50 words, explain the added benefit of using a DT regressor to this healthcare prediction problem. [2 marks]

b) With the aid of the Python packages, you will use training – test split of 80:20 to build and test two DT regression models, **Model 1 & Model 2**. The first DT model with numeric features only and the second model using all your retained features:

i. From your python notebook, insert in your report, provide as evidence the code line from your source code that ensures reproducibility of your training - test sampling. [1 marks]

ii. Using python packages, show from your code output the dimensions of your training and test subsets used for each model. List the subset of features names used for Model 1 and Model 2. [4 marks]

### Task (5) – Evaluation: How good are your models

[Total 16 Marks]

Your healthcare professionals provided the following success criteria to guide you when evaluating your models.

“When evaluating both models’ performances which addresses your research question (b), the model is expected to make some errors in estimating the survival months. However, the selected model out of the two built models should have input features that are better at explaining the recorded values of survival months.”

a) Four different regression evaluation metrics are noted. State which evaluation metric/metrics to USE or NOT USE to interpret the above success criteria closely. Justify your choice of USE or DO NOT USE. With the aid of Python packages, document the TEST SCORES for each built model. [6 marks]

Metrics	USE or DO NOT USE	Justification in relation to the success criteria	Model Name	Test Score
MSE			DT (Numeric Features Only)	
			DT (All Features Only)	
MAE			DT (Numeric Features Only)	
			DT (All Features Only)	
R-Square			DT (Numeric Features Only)	
			DT (All Features Only)	

b) Describe any caveats to your selected performance metric assessing the ability of your model meeting the success criteria. [2 marks]

c) Suggest a **single best regression model** (Model 1 or Model 2) based on the ‘USED’ performance metrics scores you identified in (Task 5. b). Briefly describe how well your selected best model satisfies the needs of the healthcare professionals. [2 marks]

d) Health care professionals aim to explain your best model’s decision of estimated survival months to the patient. Therefore, rebuild your best model while performing pre-pruning (4 levels limit) to ease the interpretation of your best model’s decision. Plot the pruned tree and paste it here from your python notebook results. Describe with evidence if there were any performance advantages or disadvantages of pruning your best tree model. [4 marks]

e) Using your pruned model, predict the survival months for breast cancer patient **B002565** whose attributes values are the following: [2 marks]

Variable Name	Value
Patient ID	<b>B002565</b>
Month of Birth	<b>July</b>
Age	<b>56 Years old</b>
Sex	<b>Female</b>
Race	<b>White</b>
Marital Status	<b>Single</b>
Occupation Code	<b>15</b>
Adopted Status	<b>Not Adopted</b>
T Stage	<b>T3</b>
N Stage	<b>N3</b>
6th Stage	<b>IIIC</b>
Differentiate	<b>Moderately differentiated</b>
Grade	<b>2</b>
A Stage	<b>Regional</b>
Tumour Size	<b>41</b>
Estrogen Status	<b>Positive</b>
Progesterone Status	<b>Positive</b>
Regional Node Examined	<b>5</b>
Regional Node Positive	<b>1</b>

---

## END OF COURSEWORK TASKS

---

### Critical notes about your coursework submission

- 1- Critical: Do not share/show your code, report, or results with any other student for any reason. Past students who shared their work with others were investigated for collusion by an Academic Misconduct Panel and awarded a zero mark.
- 2- Submit your Python notebook using the given submission link for the Python code; ensure it is in **ipynb** format. Failing to submit your Python notebook will result in a zero mark for the coursework. Python code will be used to verify your coursework outputs; discrepancies between the report and notebook may result in marks deductions/removals.

- 3- This coursework is limited to a maximum of 16 pages. The minimum font is Arial size 10 single-spaced. A minimum of 1-inch page margins. Exceeding the 16-page limit or not complying with the specified font size can result in an automatic 10% penalty deduction of your report's mark.
- 4- Use the question numbers as headers; answer the tasks in the correct order. You do not need to copy the full question; you may summarise a new header from the question, but that is unimportant. Your answers must map to each question's number and task in the correct order. Otherwise, this may lead to a significant delay in marking your work and the potential of missing out on marks lost between the lines.
- 5- There is no need to go on a new venture with coding in Python! Follow the process of code reuse. For those new to Python, all the Python code you need is given in your tutorial documents and solution Python notebooks. You only need to stitch it together from different tutorials to get the required outputs. However, I won't stop you from going on a venture with new Python coding.
- 6- Some of the submissions may be invited for a 20-minute viva. So be prepared to explain your findings should you have been invited for one. Failing to attend the viva may impact your mark.