**University of Westminster** School of Computer Science

|  |  |
| --- | --- |
| **7BUIS008W Data Mining & Machine Learning** | |
| Module leader | Mahmoud Aldraimli |
| Unit | Coursework 1 |
| 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: | This assignment contributes towards the following Learning Outcomes (LOs):   * LO2 fully implement data mining/machine learning projects, focused on problem analysis, data pre-processing, data post-processing by choosing and implementing appropriate algorithms; * LO4 fully implement encode and test data mining and machine learning algorithms using the programming language (such as Python) and standard packages and toolkits (such as R); * LO6 perform a critical evaluation of performance metrics for data mining and machine learning algorithms for a given domain/application. |
| Handed Out: | 21st February 2023 |
| Due Date | 29th March 2023  Submission by 13:00 hours |
| Expected deliverables | Submit your word template with the results/analysis on Blackboard in a zip file containing the required implemented codes in python notebook format. |
| 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, week starting 17th April 2023 |
| BCS CRITERIA MEETING IN THIS ASSIGNMENT | * 7.1.6 Use appropriate processes * 7.1.7 Investigate and define a problem * 7.1.8 Apply principles of supporting disciplines * 8.1.1 Systematic understanding of knowledge of the domain with depth in particular areas * 8.1.2 Comprehensive understanding of essential principles and practices * 8.2.2 Tackling a significant technical problem * 10.1.2 Comprehensive understanding of the scientific techniques |

Assessment regulations

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

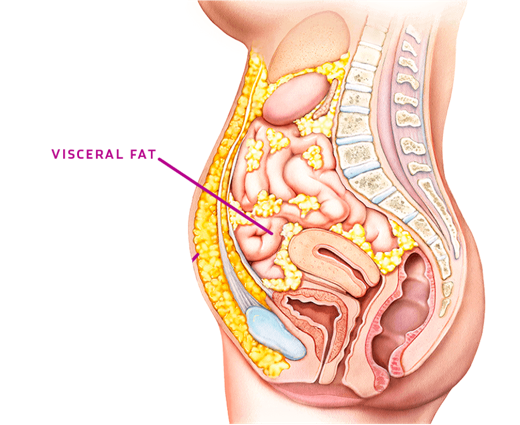
If you submit your coursework late but within 24 hours or one working day of the specified deadline, 10 marks will be deducted from the final mark, as a penalty for late submission, except for work which obtains a mark in the range 50 – 59%, in which case the mark will be capped at the pass mark (50%). 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 that 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/academic-regulations>

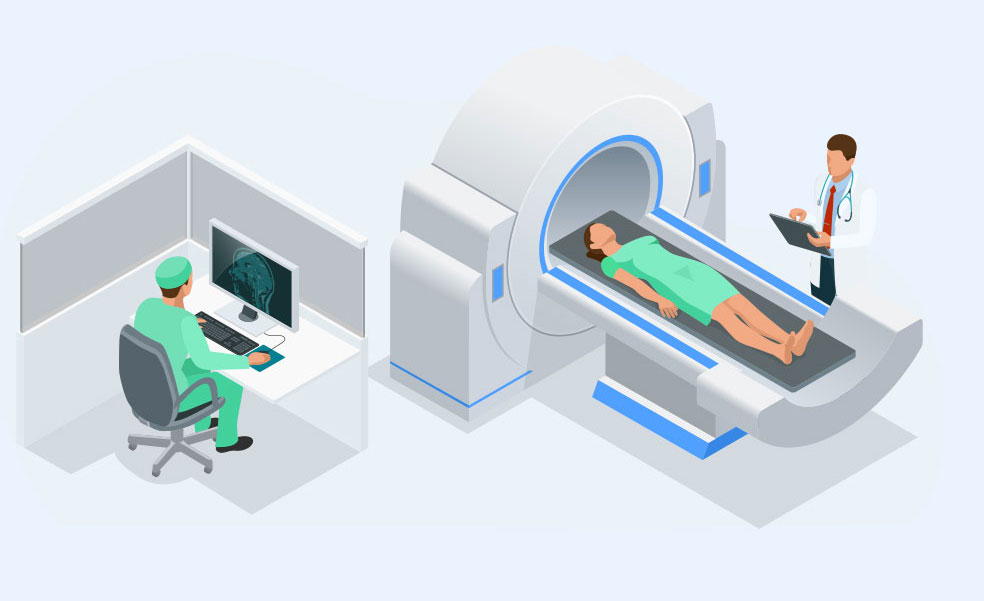
**Coursework Description**

The Real-world Problem Description

A) The Domain:

The deployment of machine learning modelling in this coursework aims at tackling a long-term *real-world* disease burden; Obesity affects an increasing number of adults in the UK, with obesity-associated changes in Adipose Tissue, else known as visceral fat (figure 1), in particular, the accumulation of visceral fat (VF) is a critical factor in determining susceptibility to diseases such as diabetes, cancers, and cardiovascular diseases. Certain levels of Visceral Fat accumulation around organs increase a person’s risk of developing chronic illnesses. Fig.1 Illustration of visceral fat (VF)

Suppose individuals with excessive amounts of visceral fat are identified. In that case, suitable diet, exercise and medical interventions can be recommended/prescribed to reduce such fat in an effort to reduce their risk of developing a chronic disease.



B) The Current Process of identifying VF amounts:

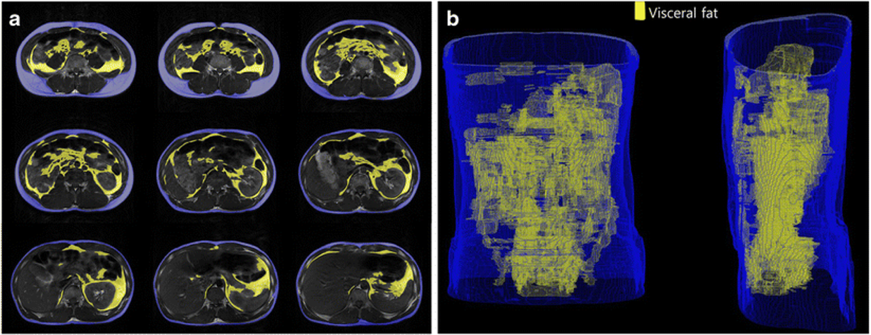
Visceral fat (VF) amounts in the human body are measured in units of litres. The method currently used to measure the amount of VF within the body is achieved by viewing the inside of the body via Magnetic Resonant Imaging (MRI) scan (see Figure 2). The result of the scan is multiple images that contain regions of interest. These regions are examined by a radiologist to indicate the location and amounts of visceral fat (VF) in the body, see Figure 3.

Fig.2 Illustration of MRI body scan process

Fig.3 Illustration of MRI body scan images

C) The Domain Problem:

The MRI scan cannot be done for everyone. The MRI scan process is expensive; without the NHS and the taxpayers' money, each MRI scan can cost an average of £450 to £600. Also, the MRI scan is time-consuming. A single scan can take up to 90 minutes to complete. The MRI images need to be examined by a radiologist to distinguish those with dangerous VF amounts; the results can take a couple of weeks to arrive. Due to cost, it is not feasible to do an MRI for everybody to establish those at risk of developing long-term critical illnesses. Also, VF cannot be predicted by the naked eye, some people are thin looking from the outside (slim), but they still could carry dangerous levels of VF inside (Thin Outside and Fat Outside). Again, some athletes, like sumo wrestlers, look thick on the outside (Obese), but they have low VF amounts; they are known as Thin Inside and Fat outside. See Figure 4.

Fig.4 Illustration of Fat Outside Fat Inside and Thin Outside Fat Inside

From research, depending on VF amounts in litres, subjects’ risk of developing diseases can be either Low-Risk, Moderate-Risk or High-Risk. Unlike Low Risk, Moderate and High-Risk subjects require changes (interventions) to their lifestyle and diet. Unlike Moderate Risk subjects, High-Risk Subjects require urgent, or even immediate changes to lifestyle and diet.

Additional domain challenge is also linked to the physiological differences between males and females; dangerous VF levels in females differ from males. See Table 1 below.

Table 1. The definition of levels of VF in both females and males

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gender | VF Levels | Risk Level | Require Intervention (Y/N)? | Intervention Urgency |
| Females | VF ≤2 L | Low Risk | No | None |
| 2< VF ≤5 L | Moderate Risk | Yes | Medium |
| VF > 5 L | High Risk | High |
| Males | VF ≤3 L | Low Risk | No | None |
| 3< VF ≤6 L | Moderate Risk | Yes | Medium |
| VF > 6 L | High Risk | High |

D) Your Role as A Data Scientist:

You are hired as a data scientist to work alongside a team of healthcare professionals to build machine learning models to tackle the problem of above problem of Visceral Fat obesity. The healthcare professionals have two historical datasets, one for males and the other for females, who underwent MRI in the past and recorded their VF levels in litres.

The Healthcare professionals are relying on your help to answer only one of the following three questions on just one of the two datasets:

1. Does machine learning have the potential to predict the risk levels (Low, Moderate, High) of Visceral Fat for new future subjects based on historical Visceral Fat data without the need for MRI scans in the future?
2. Does machine learning have the potential to predict new future subjects who need lifestyle and diet interventions (Intervention Required vs Not Required) based on the historical Visceral Fat data without the need for MRI scans in the future?
3. Does machine learning have the potential to predict the urgency of intervention for subjects who need lifestyle and diet changes (None, Medium, High) based on their Visceral Fat without the need to undergo MRI scans in the future?

E) Your Dataset

You are given two datasets, one for females and the other for males. As mentioned, due to physiological differences, one must not mix both cohorts, they are analysed independently. The dataset contains the following attributes:

|  |  |  |
| --- | --- | --- |
| Attribute Name | Attribute Description | Unit |
| SUBJECT\_ID | Unique alphanumeric identification number per subject | None |
| SEX | Subject’s gender M: Male and F: Females | None |
| AGE | Subject’s age. | Years |
| BMI | Body Mass Index. Calculated by: Weight(Kg)/[Height(m)]2 | Kg/m2 |
| HEIGHT | Subject’s measured height on the scan day | cm |
| WEIGHT | Subject’s measured weight on the scan day | kg |
| WAIST\_CIRCUMFERENCE | A length measurement around the middle of the body | cm |
| DIASTOLIC\_BLOOD\_PRESSURE | The pressure in arteries when heart rests between beats | mmHg |
| SYSTOLIC\_BLOOD\_PRESSURE | The pressure in arteries when heart beats | mmHg |
| WALK\_DURATION\_PER\_DAY | Duration of subject’s walk in a single 24 hours (one day) | Minutes |
| COMPUTER\_USE\_TIME\_PER\_DAY | Duration of subject’s PC use in a single 24 hours (one day) | Hours |
| SMOKING\_STATUS | 1: Smokers and 0: Non-smoker | None |
| DISCONTINUED\_NO | Unknown | None |
| CIGARETTES\_CONSUMED\_PER\_DAY | Count of cigarettes smoked in a single 24 hours (one day) | Count |
| Visceral\_Fat\_Volume | Visceral fat volume calculated from MRI scan | Litres |

The Machine Learning for Classification Coursework Tasks

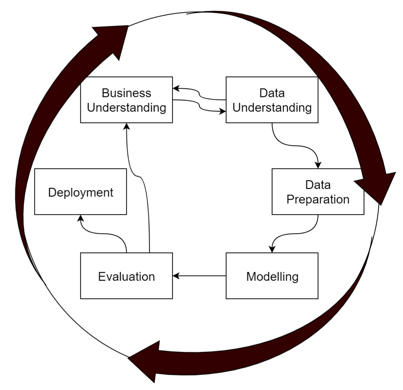
As a data scientist, you are a logician, a mathematician, a technician, and an analyst, and you need healthcare professionals to understand your analysis. Healthcare professionals are busy individuals, and they don’t have all the time in the world. One essential skill that you must adhere to is to be straight to the point. Focus on the answer needed for each task, and provide 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 the results, so you must not paste any python code in this report unless you are specifically asked to. You will be provided with 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 except for deployment (See figure 5)

Fig.5 CRISP-DM Phases

Task (1) – Domain Understanding: Classification or Regression [Total 1 Mark]

1. It was decided by your healthcare team that this problem is a classification problem. Logically, describe in one sentence or two a potential reason for proceeding with classification predictive modelling instead of regression predictive modelling to answer any one of the healthcare professionals' questions. [1 Mark]

Task (2) – Data Understanding: Producing Your Experimental Designing [Total 7 Marks]

1. You are required to note down which gender dataset you want to use for modelling. Choose only one dataset. Note down your selected single healthcare question that you are attempting to answer in this coursework and the type of classification problem you have. [2 Marks]
2. From your python notebook, show a data frame of the first 10 records of the possible input variables and your class variable. Produce a statistical description and measurement scale type of your selected dataset attributes. Emphasize the distribution of your class variable. (Use screenshots of code outputs only). [5 Marks]

Task (3) – Data Preparation: Cleaning and Transforming your data [Total 30 Marks]

1. Investigate any found issues in your selected dataset and the possible variables. Use the table below to organise your findings: [10 Marks]

|  |  |  |
| --- | --- | --- |
| Dataset or Variable | Name of variable | Issue description |
| i.e., Variable | i.e., Duration of sleep | The issue with this variable is ………… |
| i.e., Variable | i.e., Job title | The issue with this variable is ………… |
| i.e., Whole Dataset | Whole dataset | The issue with this variable is ………… |
|  |  |  |

1. Based on the issues you found in your data, suggest a suitable possible solution to mitigate each of these issues and provide your justification for using each solution. Again, organise your answer in a table as before. See below: [10 Marks]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset or Variable | Name of variable | The Issue | Solution | Justification |
| i.e., Variable |  |  |  |  |
| i.e., Variable |  |  |  |  |
| i.e., Whole Dataset |  |  |  |  |
|  |  |  |  |  |

1. With the aid of Python packages and a notebook, implement your suggested solutions, and show evidence of implementing your suggested solutions to the problems you identified for your dataset in (a) and (b). (Use screenshots of code outputs only). Indicate which issue was resolved from each screenshot provided. Show screenshots of code outputs before and after implementing your solution. [10 Marks]

Task (4) – Modelling: Create Predictive Classification Models [Total 15 Marks]

1. From the classification algorithms which you learnt in the module, four different algorithms were selected, Naïve Bayes (NB), Decision Trees (DT), K-Nearest Neighbour (KNN) and Artificial Neural Networks (ANN) with a Multi-Layer Perceptron Architecture. These algorithms are a mix of parametric and non-parametric algorithms. List down the type of each algorithm (parametric vs non-parametric), name any hyperparameters for each algorithm which you may want to consider tuning, and the python package (python source code) used for calling each algorithm. Again, organise your answer in a table as before. See below:

[12 marks]

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm Name | Type of Algorithm | Possible Hyper -parameters | Python package source code to call the algorithm |
| NB |  |  |  |
| DT |  |  |  |
| KNN |  |  |  |
| ANN(MLP) |  |  |  |

1. With the aid of the Python packages, using the training–test split approach, use the numeric input features only and the output feature, to build your predictive classification models. Screenshot the head of the data frame used for training the algorithms. In a just few sentences, justify your choice of the training-test split ratio, and provide a text reference if necessary. Provide as evidence the code line from your source code that guarantees that all models were tested on the same test dataset.

[3 Marks]

Task (5) – Evaluation: How good are your models [Total 47 Marks]

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

“When evaluating your models' performance, which addresses your selected question out of the three questions, the first priority is identifying the largest number of subjects related to the high and moderate-risk groups to recommend interventions and potentially save lives. But you need to keep in mind, and it also needs to reasonably identify those who are related to low-risk, to minimise any unnecessary initiation of lifestyle and diet changes.”

1. With the aid of python packages, paste the test confusion matrix for each trained model as screenshots from the output of your python code. [4 marks]
2. With the aid of python packages, five different classification evaluation metrics are noted. State which of them you think are strongly related (interpret) to the above success criteria and which are not. Provide a justification for related or unrelated. Document the score for each built model. [30 marks]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric Name | Related or Unrelated | Justification in relation to the success criteria | Model Name | Metric Score |
| Accuracy |  |  | NB |  |
| DT |  |
| KNN |  |
| ANN |  |
| Recall |  |  | NB |  |
| DT |  |
| KNN |  |
| ANN |  |
| Precision |  |  | NB |  |
| DT |  |
| KNN |  |
| ANN |  |
| F-Measure |  |  | NB |  |
| DT |  |
| KNN |  |
| ANN |  |
| AUC-ROC |  |  | NB |  |
| DT |  |
| KNN |  |
| ANN |  |

1. Based on the related performance metrics scores you identified in (b), suggest the best classification model. Briefly describe how this model satisfies the needs of your healthcare professionals. [3 marks]
2. To enhance your best model’s generalisation performance, you can tune its hyperparameters which you indicated in (a) for that specific algorithm. With the aid of python packages, Re-train the algorithm again with GridSearchCV, and indicate the number of cross-validation K folds used. For the newly tuned model, document the estimated best hyperparameters, present the test confusion matrix and calculate and document the new scores of the related metrics to the success criteria identified in (b). Explain your observations on whether the tuning of hyperparameters enhanced the generalisation of your original best model.

[6 Marks]

1. Based on your best model, draft an answer for your selected research question, provide criticism of 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 200 words. [4 Marks]

Critical note about the structure of your submission:

1. This coursework is limited to a maximum of 7 pages. The minimum font is Arial size 12 single-spaced. A minimum of 1-inch page margins. Exceeding the page limit or going below the specified font size will result in an automatic 20% penalty deduction of your report’s mark.
2. Use the question numbers as headers; you do not need to copy the full question, you may summarise a new header from the question, but that is not important. It is crucial that your answers map to each question’s number and 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.
3. There is no need to go on a new venture with coding in python! Follow the process of process code reuse. For those who are new to python, all the python code that you need is given in your tutorial documents and solution python notebooks. All you need is 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.
4. 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.