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:	25th October 2023
Due Date	30th November 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, the week starting 8th January 2024
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 new *real-world pandemic*. The disease mpox (formerly monkeypox) is caused by the monkeypox virus (commonly abbreviated as MPXV), an enveloped double-stranded DNA virus of the *Orthopoxvirus* genus in the *Poxviridae* family, which includes variola, cowpox, vaccinia and other viruses. The two genetic clades of the virus are clades I and II.

The monkeypox virus was discovered in Denmark (1958) in monkeys kept for research, and the first reported human case of mpox was a nine-month-old boy in the Democratic Republic of the Congo (DRC, 1970). Mpox can spread from person to person or occasionally from animals to people. Following the eradication of smallpox in 1980 and the end of smallpox vaccination worldwide, mpox steadily emerged in central, east and west Africa. A global outbreak occurred in 2022–2023. The natural reservoir of the virus is unknown – various small mammals such as squirrels and monkeys are susceptible.

In May 2022, an outbreak of mpox appeared suddenly and rapidly spread across Europe, the Americas and then all six WHO regions, with 110 countries reporting about 87 thousand cases and 112 deaths. The global outbreak has affected primarily (but not only) gay, bisexual, and other men who have sex with men and has spread person-to-person through sexual networks. In 2022, outbreaks of mpox due to Clade I MPXV occurred in refugee camps in the Republic of Sudan. A zoonotic origin has not been found.

Surveillance, diagnostics, risk communication and community engagement, remain central to stopping the outbreak and eliminating human-to-human transmission of mpox in all contexts.

Transmission

Person-to-person transmission of mpox can occur through direct contact with infectious skin or other lesions, this includes contact which is

- face-to-face (talking or breathing)
- skin-to-skin contact
- mouth-to-mouth
- mouth-to-skin contact
- respiratory droplets or short-range aerosols from prolonged close contact

The virus then enters the body through broken skin, mucosal surfaces (e.g., oral, pharyngeal, ocular, genital, anorectal), or via the respiratory tract. Mpox can spread to other members of the household and to sex partners. People with multiple sexual partners are at higher risk.

Animal-to-human transmission of mpox occurs from infected animals to humans from bites or scratches or during activities. The extent of viral circulation in animal populations is not entirely known, and further studies are underway.

People can contract mpox from contaminated objects such as clothing or linens, through sharps injuries in health care, or in a community setting such as tattoo parlours.

B) The current process of diagnoses Monkeypox:

Identifying mpox can be difficult as other infections and conditions can look similar. It is important to distinguish mpox from chickenpox, measles, bacterial skin infections, scabies, herpes, syphilis, other sexually transmissible infections, and medication-associated allergies. Someone with mpox may also have another sexually transmissible infection, such as herpes. Alternatively, a child with suspected mpox may also have chickenpox. For these reasons, testing is key for people to get treatment as early as possible and prevent further spread.

Detection of viral DNA by polymerase chain reaction (PCR) is the recommended laboratory test for mpox. The best diagnostic specimens are taken directly from the rash – skin, fluid or crusts – collected by vigorous swabbing. Without skin lesions, testing can be done on oropharyngeal, anal or rectal swabs. See Figure1.

Fig.1 MPOX PCR Home Testkit

C) The Domain Problem:

There are different types of MPOX tests: polymerase chain reaction (PCR) tests. PCR tests are used to directly screen for the presence of viral RNA, which will be detectable in the body before antibodies form or symptoms of the disease are present. This means the tests can tell if someone has MPOX early in their illness.

These types of MPOX tests need to be sent away to a laboratory for analysis, meaning it can take days for people to find out their results. Most tests cost around £100. Most private clinics charge around £200 for the tests.

It is 2026, and the World Health Organisation (WHO) announced a new variant of the MPOX outbreak and declared a pandemic. After the return from a 12 month-lockdown, all workforce in the UK is required to do a bi-weekly PCR test. Due to demand, the PCR test cost is higher than before.

D) Your Role as A Data Scientist:

You are hired as a data scientist to work alongside a team of healthcare professionals and virologists to build machine learning models to tackle the above problem of MPOX screening. The healthcare professionals have a historical dataset of those who underwent a PCR test in 2022 and recorded their results.

The Healthcare professionals are relying on your help to answer the following **research question** on the dataset; the key requirement is to create a new, inexpensive screening tool that does not require lab results:

1) Does machine learning have the potential to create an inexpensive screening tool to predict those who could have contracted the MPOX virus without undertaking a PCR test to minimise the number of required PCR tests performed by the workforce population?

E) Your Dataset

The dataset contains the following attributes:

Attribute Name	Attribute Description	Unit
Test ID	The identification number of the PCR test sample	None
Systemic Illness	Symptoms affecting the entire body (multiple values)	None
Sore Throat	Occurrence of Sore Throat – (True, False)	None
Rectal Pain	Occurrence of Rectal Pain– (True, False)	None
Penile Oedema	Occurrence of Penile Oedema– (True, False)	None
Oral Lesions	Occurrence of Oral Lesions– (True, False)	None
Solitary Lesion	Occurrence of Solitary Lesion– (True, False)	None
Swollen Tonsils	Occurrence of Swollen Tonsils– (True, False)	None
HIV Infection	Status of HIV Infection– (True, False)	None
Red blood cells	Red blood cells count	cells/mcL
White blood cells	White blood cells count	cells/mcL
Home ownership	The patient's home ownership status (1:Yes, 0:No)	None
Age	The patient's age at the test	Years
Month of Birth	The patient's month of birth	None
Health Insurance	Having Private Health Insurance (1: Yes, 0: No)	None
Sexually Transmitted Infection	History of Sexually Transmitted Infection	None
MPOX	The PCR Test Result	None

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 specifically asked to. 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 except for deployment (See Figure 5)

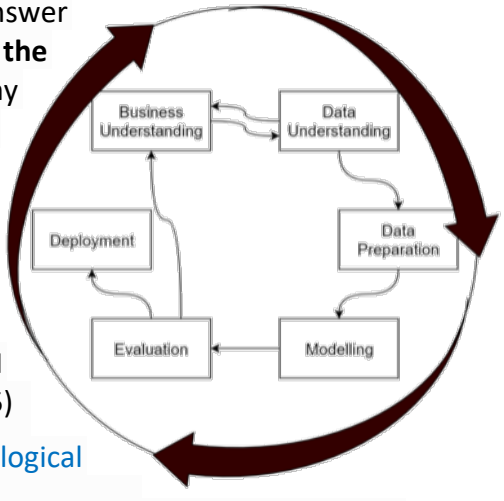

Fig.5 CRISP-DM Phases

Note: You must answer each task in the given chronological order and use the questions as headers.

Task (1) – Domain Understanding: Classification [Total 4 Mark]

a) The healthcare professionals decided that only 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 (logically applicable) in the classification modelling of MPOX classes and those you should DROP (REMOVE). [4 Mark]

Attribute Name	Retain or Drop
Test ID	
Systemic Illness	
Sore Throat	
Rectal Pain	
Penile Oedema	
Oral Lesions	
Solitary Lesion	
Swollen Tonsils	
HIV Infection	
Red blood cells	
White blood cells	
Home ownership	
Age	
Month of Birth	
Health Insurance	
Sexually Transmitted Infection	
MPOX	

Task (2) – Data Understanding: Producing Your Experimental Designing [Total 4 Marks]

a) From your Python notebook, for the **RETAINED input variables and your class output variable, produce a basic statistical description** and **measurement scale type** of your retained attributes. Plot the **distribution of your class variable**. (Use screenshots of code outputs only). [4 Marks]

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

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

Dataset or Variable Issue?	Name of variable	Issue description
i.e., Variable Issue		The issue with this variable is ……
i.e., Variable Issue		The issue with this variable is ……
i.e., Variable Issue		The issue with this variable is ……
i.e., Whole Dataset	Whole dataset	The issue with this dataset is ……
⋮	⋮	⋮

b) 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 Issue?	Name of variable	The Issue	Solution	Justification
i.e., Variable Issue				
i.e., Variable Issue				
i.e., Variable Issue				
i.e., Whole Dataset	Whole dataset			
⋮	⋮	⋮	⋮	⋮

c) 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]

a) From the classification algorithms which you learned in the module, four different algorithms were selected: Naïve Bayes (NB), Decision Tree (DT), Logistic Regression (LR) and Support Vector Machine (SVM) with an RBF kernel. 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 hyperparameters for each algorithm which you may want to consider tuning. Note down 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	Learnable Parameters	Possible Hyper - Parameters	Python package source code to call the algorithm
LR				
DT				
KNN				
SVM (RBF)				
NB				

b) With the aid of the Python packages, using the training–test split approach, use the applicable categorical input features only and the class output feature to **build your predictive classification models**. Screenshot the list of all feature names used for building the classification models and the corresponding data shape function output.

In just a few sentences, research and justify your choice of the **training-test split ratio** and provide an in-text reference. 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 MPOX Negatives : MPOX Positives is the same in the training and test sets**. [3 Marks]

Task (5) – Evaluation: How good are your models [Total 47 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 selected research question. The model is expected to misclassify subjects. Thus, the model should aim to predict as many subjects as MPOX positive subjects as possible to recommend performing a PCR test and self-isolation. However, the model should demonstrate that its high MPOX positive prediction rate is mainly due to a larger portion of correctly detected subjects among all predictions made belonging to the MPOX positive 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. [4 marks]

b) **Five different classification evaluation metrics** are noted. State which evaluation metric/metrics to use to strongly interpret the above success criteria (interpret the success criteria) and which are not. Justify your choice of **USE or DO NOT USE**.
With the aid of Python packages, document the **TEST SCORES** for each built model. [25 marks]

Metric Name	**"USE" or "DO NOT USE"**	**Justification in relation to the success criteria**	**Model Name**	**Metric Score**
Accuracy			LR	
			DT	
			KNN	
			SVM (RBF)	
			NB	
Recall			LR	
			DT	
			KNN	
			SVM (RBF)	
			NB	
Precision			LR	
			DT	
			KNN	
			SVM (RBF)	
			NB	

F-Measure			LR	
			DT	
			KNN	
			SVM (RBF)	
			NB	
AUC-ROC			LR	
			DT	
			KNN	
			SVM (RBF)	
			NB	

c) Based on the **'USED'** performance metrics scores you identified in (Task 5. b), suggest the **best classification model or models**. Briefly describe **how this model satisfies** the needs of your healthcare professionals. [3 marks]

d) To enhance your selected best model/s performance, you can tune its 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 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 **USED** metrics to interpret the success criteria identified in (Task 5.b). **Explain your observations** on **whether** the tuning of hyperparameters enhanced the generalisation of your original best model. [6 Marks]

e) Considering the models created in Task (4-b), combine only two learners in an ensemble voting learner. In relation to each base learner's test confusion matrix, specify your reasoning behind the choice of both base learners. Using the test confusion matrices, explain if any performance improvement is made by combining both base learners into a voting ensemble learner. [5 Marks]

f) Based on your best model, draft an answer for the 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. State any ethical issues that your model may raise if used to screen for MPOX. [4 Marks]

Critical note about the structure of your submission:

1- This coursework is limited to a maximum of 13 pages. The minimum font is Arial size 12 single-spaced. A minimum of 1-inch page margins. Exceeding the page limit or not complying with the specified font size will result in **an automatic 10% penalty deduction** of your report's mark.

2- 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. It is crucial that your answers 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.

3- There is no need to go on a new venture with coding in Python! Follow the process of process 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.

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.