THE UNIVERSITY OF HUDDERSFIELD

School of Computing and Engineering

ASSIGNMENT SPECIFICATION

Module	Details
Module Code	CHA2555
Module Title	Artificial Intelligence
Course Title/s	BSc (Hons)/MSci Computer Science, BSc (Hons)/MEng Software Engineering, BSc (Hons)/MComp Computing, BSc (Hons) Computer Science with Cyber Security, BSc (Hons) Computer Science with Games Programming, BSc (Hons) Applied Computing (Top-up)

Assessment	Weighting, Type and Contact Details
Title	Data-driven Artificial Intelligence
Weighting	50%
Mode of working for assessment task	Individual This assessment task is to be completed on an individual basis and there should be no collusion or collaboration whilst working on and subsequently submitting this assignment.
Module Leader	George Bargiannis (g.bargiannis@hud.ac.uk)
Module Tutors	Emmanouil Papadakis (e.papadakis@hud.ac.uk) Bakhtiar Amen (b.amen@hud.ac.uk)

Submission	Submission and Feedback Details
Hand-out date	Monday 31 October 2022
How to submit your work.	Brightspace submission point
Submission date/s and times	Monday 9 January 2023 23:59
Expected amount of independent time you	15 hours

Submission	Submission and Feedback Details
should allocate to complete this assessment	
Submission type and format	A written report in the form of a PDF document
Date by which your grade and feedback will be returned	Monday 30 January 2023

Additional Guidance Information	Details
Your responsibility	It is your responsibility to read and understand the University regulations regarding conduct in assessment. Please pay special attention to the assessment regulations (section 10) on Academic Misconduct. In brief: ensure that you; 1. DO NOT use the work of another student - this includes students from previous years and other institutions, as well as current students on the module. 2. DO NOT make your work available or leave insecure, for other students to view or use. 3. Any examples provided by the module tutor should be appropriately referenced, as should examples from external sources. Further guidance can be found in the SCEN Academic Skills Resource and UoH Academic Integrity Resource module in Brightspace. If you experience difficulties with this assessment or with time management, please speak to the module tutor/s, your Personal

Additional Guidance Information	Details
	Academic Tutor, or the School's Guidance Team. (sce.guidance@hud.ac.uk).
Requesting a Late Submission	It is expected that you complete your assessments by the published deadlines. However, it is recognised that there can be unexpected circumstances which may affect you being able to do so. In such circumstances, you may submit a request for an extension. Extension applications must be submitted before the published assessment deadline has passed. There are two types of extension that you may request. You will be required to indicate which one you are applying for when you submit the request for Late Submission via MyHud/MyStudies. 1. Self-certified illness extension of up to 5 working days. • Evidence will not be required for this type of request, but you are limited to two self-certified extension requests in any academic year. 2. Extension request of up to 10 working days. • This extension requires you to submit appropriate evidence in support of your request. The maximum extension that can grant is 10 working days. <u>Accepted grounds for an extension</u> • Serious short-term illness or accident (of a nature which in employment would result in a health-related absence); • Evidence of a long-term health condition worsening; Emerging mental health condition, or worsening of an existing mental health condition; • Bereavement.

Additional Guidance Information	Details
	If you are unable to submit work within the maximum late submission period of 10 days, contact the School's Guidance Team. (sce.guidance@hud.ac.uk), as you may need to submit a claim for Extenuating Circumstances (ECs).
Extenuating Circumstances (ECs)	An EC claim is appropriate in exceptional circumstances, when an extension is not sufficient due to the nature of the request. You can access the EC claim form on the Registry website; where you can also find out more about the process. You will need to submit independent, verifiable evidence for your claim to be considered. Once your EC claim has been reviewed you will get an EC outcome email from Registry. If you are unsure what it means or what you need to do next, please speak to the Student Support Office – Room SJ1/01 An approved EC will extend the submission date to the next assessment period (e.g July resit period).
Late Submission (No ECs approved)	Late submission, up to 5 working days, of the assessment submission deadline, will result in your grade being capped to a maximum of a pass mark. Submission after this period, without an approved extension, will result in a 0% grade for this assessment component.
Tutor Referral available	YES
Resources	• Please note: you can access free Office365 software and you have 1 Tb of free storage space available on Microsoft's OneDrive – Guidance on downloading Office 365.

Assignment 2: Data-driven Artificial Intelligence

1. Assignment Aims

- To develop understanding of data-driven AI concepts and methods.
- To gain experience in solving problems using data-driven AI and machine learning methods.

2. Learning Outcomes:

- Explain some of the main sub-symbolic approaches used to implement intelligent systems, such as neural networks or Bayesian networks.
- Describe some of the central techniques in specific AI areas e.g. knowledge representation and/or machine learning.
- Configure, apply and critically evaluate machine learning methods, and appropriate tools and techniques, for implementing intelligent systems in application areas.

3. Assessment Brief

This assignment consists of 8 Questions. Question 1 is on AI planning, while Question 2 focuses on concepts and methods related to data-driven AI. Questions 3-8 involve applying several data-driven AI and machine learning methods on provided data sets.

Question 1 [15%]

Consider the following states:

- Initial state: At(Home)
- Goal state: At(Home), Travelled

and the following actions:

- Action: AirportCheckIn
    - Precondition: At(Airport)
    - Effect: Have(BoardingPass), ~Have(Luggage)

- Action: FlightCancelled
    - Precondition: At(Airport), Have(BoardingPass)
    - Effect: ~Checked, ~Have(BoardingPass)
- Action: GetLuggage
    - Precondition: At(Home)
    - Effect: Have(Luggage)
- Action: GoHome
    - Precondition: At(Airport), Travelled
    - Effect: ~At(Airport), At(Home), ~Checked, ~Have(BoardingPass)
- Action: GoToAirport
    - Precondition: At(Home)
    - Effect: At(Airport), ~At(Home)
- Action: SecurityCheck
    - Precondition: At(Airport), Have(BoardingPass)
    - Effect: Checked
- Action: RoundTrip
    - Precondition: At(Airport), Checked, Have(BoardingPass)
    - Effect: ~Checked, ~Have(BoardingPass), Travelled

Starting from the initial state, generate a state diagram listing all the possible states and possible transitions.

Question 2 [10%]

For each of the following multiple-choice questions, choose those sentences that are correct. *Note that you will only be awarded points if you choose all correct sentences.*

1. Machine Learning (ML) and Artificial Intelligence (AI)
    a. Machine Learning is a type of Artificial Intelligence
    b. Machine Learning can solve any problem as soon as data is available.
    c. Choosing an appropriate machine learning technique depends on the problem.
    d. A machine learning application generates patterns of data from existing knowledge.
2. Machine Learning algorithms
    a. Clustering is a type of primitive classification
    b. Unsupervised learning requires a sample of labelled data
    c. Regression is used for the prediction of nominal values given a range of nominal and/or numerical values.
    d. Unsupervised learning allows the discovery of hidden patterns in the data.
3. Machine learning process
    a. Data cleaning is an optional task in a machine learning pipeline
    b. An outlier refers to a mislabelled example or error in the data.

c. Noise in the data refers to data entries that do not “fit” with the rest of the data set.
d. Every Machine Learning algorithm requires its own unique tuning (parameter selection).
e. Using the same data for learning and testing causes the mode to underfit, which means the model achieves a perfect score but fails to generalise on yet-unseen data.

Question 3 [10%]

With the given data set waterQuality.csv, you need to show hands-on skills of using Weka toolkit to perform classification tasks with the J48 Decision Tree.

In particular:

1. Apply 10-fold cross validation and classify water quality based on the safety for consumption.
2. Visualise the accuracy and the generated decision tree.
3. Inspect the attributes to identify impurities (missing values, outliers, noise, imbalance etc).
4. Process and clean the data accordingly.
5. Re-apply classification and assess the new performance and compare it with step 1.
6. Apply feature selection (using 10-fold validation)
7. Remove any “weak” features (threshold < 70%) from the data set.
8. Reclassify and assess the new performance and compare it with step 5.

In your report you need to include a brief summary of your data cleaning work for steps 1-8, including a list of identified impurities in step 3 and the following 6 screenshots:

- Screenshot for every classification run, with the confusion matrix and performance visible (3 in total).
- Screenshot of the outliers (distribution) before and after cleaning the data (at least 1).
- One screenshot of the feature selection result.
- One screenshot of the final decision tree structure.

Question 4 [20%]

Consider the following database of stars represented by 5 training examples. The target attribute is ‘Red dwarf’, which can have values ‘yes’ or ‘no’. This is to be predicted based on the other attributes of the star. Note that Radius refers to the relative radius based on the Sun of our solar system and the temperature is measured in Kelvin.

Star	Temperature	Radius	Colour	Red dwarf
1	3650	1324	Red	No
2	8930	0.0095	White	No
3	3511	0.109	Red	Yes
4	3570	1480	Red	No
5	2840	0.11	Red	Yes

1. Calculate the entropy of the target attribute. Recall that $Entropy(S) \stackrel{\text{def}}{=} \sum_{i=1}^{c} p_i log_2\left(\frac{1}{p_i}\right) = \sum_{i=1}^{c} -p_i log_2(p_i)$, where p_i is the proportion of S belonging to class i. Also note that $\log_2(x) = \frac{\log_{10}(x)}{\log_{10}(2)}$.
2. Construct the decision tree structure from the above examples, which would be learned by the ID3 algorithm.
    - For your convenience, convert the numerical values to categorical by comparing each star with the properties of the Sun:
        - Temperature > 6000K → Hotter
        - Temperature < 6000K → Cooler
        - Radius > 1 → Larger
        - Radius < 1 → Smaller
3. Show the value of the information gain for each candidate attribute at each step in the construction of the tree.

In your report you need to include all details of the tree construction process following steps 1-3.

Question 5 [5%]

Given the clean data set produced in step 4 of Question 3, use Weka to perform classification tasks with multiple algorithms and estimate the best performance. In this question you will need to:

1. Apply a feature selection process (information gain), using **bacteria** as the class attribute.
    - Note: the attribute bacteria contains numerical values, this needs to be converted into a nominal data type (e.g., low, moderate, high).
2. Remove the "weak" features (threshold < 70%) from the data set.

3. Compare the performance of the models: **REPTree** and **Linear Regression** while predicting the population of bacteria per litre of water. (*use 10-fold validation and default settings).*
4. Choose the best model.
5. Compare how the performance of the optimal model changes when different data sets are used for validation: a) training set, b) 10-fold validation and c) 50/50 data set split (50% training and 50% testing).

In your report you need to include 2 screenshots that depict the performance of each classification model, and answers to the following questions:

- What is the best regression model?
- Explain why the best model outperforms the other.
- What is the performance of the best model using different validation methods, which is the most reliable and why?

Question 6 [12%]

Consider a medical diagnosis problem in which there are two alternative hypotheses: (1) that the patient has a particular form of cancer, and (2) that the patient does not. The available data is from a particular laboratory test with two possible outcomes: + (positive) and - (negative). We have prior knowledge that only 0.7% over the entire population have this disease. Furthermore, the lab test is only an imperfect indicator of the disease. The test returns a correct positive result in 95% of the cases in which the disease is actually present and a correct negative result in 96% of the cases in which the disease is not present. In other cases, the test returns the opposite result.

1. Summarise the above problem description using probabilities.
2. Suppose we now observe a new patient for whom the lab test returns a positive result. What is the probability of the patient being with/without cancer?

Question 7 [10%]

Given the clean data set produced in step 4 of Question 3, use Weka to perform unsupervised learning tasks with clustering algorithms. The goal is to conduct an explanatory analysis by clustering the given data set, in order to discover which among the chemicals found in water have the most significant impact on the population of viruses.

In particular, you need to:

1. Clean the data from irrelevant attributes for the goal of exploratory analysis.
2. Use SimpleKMeans on the training set, which is a Weka implementation of the conventional k-means algorithm to apply clustering.

3. Experiment with different numbers of clusters, keeping the other parameters settings as default
4. Use the Weka-generated cluster assignments tool to visualise the most "useful" clusters for the given goal

In your report you need to include:

- The list of attributes that will be used in the clustering procedure.
- The optimal number of clusters.
- At least 3 screenshots of the learnt centroids for different number of clusters (**must include the learnt centroids for the optimal number of clusters**).
- At least 3 screenshots that depict the findings of the explanatory analysis, along with the necessary justification.

Question 8 [18%]

Use the k-means algorithm and Euclidean distance to cluster the following 8 examples into 3 clusters: A1=(4,10), A2=(5,8), A3=(1,7), A4=(4,7), A5=(6,5), A6=(6,10), A7=(8,5), A8=(4,9).

Suppose that the initial seeds (centres of each cluster) are A3, A5 and A7. Run the k-means algorithm for 1 epoch only.

In particular:

1. Fill the distance matrix based on the Euclidean distance of the points given above:

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0							
A2		0						
A3			0					
A4				0				
A5					0			
A6						0		
A7							0	
A8								0

2. Calculate the cluster assignment at the end of the first epoch:
    a. The new cluster assignment (i.e. contents of each cluster)
    b. The centroids of the new clusters
3. How many more iterations are needed to converge? Show cluster assignments and updated centroids for each of the remaining epochs.

In your report, you need to include the appropriate cluster assignment and centroids.

4. Marking Scheme

Marking Criteria for Question 1 [8 + 7 = 15%]

a) Up to 8 points can be given based on how many states are identified correctly
b) Up to 7 points can be given based on how many actions have been applied to the correct states

Marking Criteria for Question 2 [3+3+4= 10%]

Points are given on the basis of the correct choices made.

Marking Criteria for Question 3 [6 + 4 = 10%]

The screenshots should be able to reflect what's asked in the questions, and you must explicitly identify what the screenshots correspond to in order to get full marks. Also, the written answers would be marked based on completeness. Specifically, marking points: Screenshots for classification runs (3%), feature selection (2%), final tree structure (1%) and balancing (1%). Written answers for data cleaning (3%).

Marking Criteria for Question 4 (3 + 7 + 10 = 20%)

a) 3 points will be given for the correct entropy value
b) Up to 7 points can be given based on how much information is added on the drawn decision tree.
c) Up to 10 points can be awarded based on the number of information gains given .

Marking Criteria for Question 5 (3+2=5%)

The two screenshots should be able to reflect what is asked in the questions, and you must explicitly identify what the screenshots correspond to in order to get full marks. Also, the written answers would be marked based on completeness. Specifically, marking points: Screenshots for classification runs (1%), written answer 1 (1%) and written answer 2 (3%).

Marking Criteria for Question 6 (6 + 6 = 12%)

a) Up to 6 points can be given based on the number of correct given probabilities
b) Up to 6 points can be given if you supply the correct final answers.

Marking Criteria for Question 7 (6 + 4 = 10%)

The two screenshots should be able to reflect what's asked in the questions, and you must explicitly identify what the screenshots correspond to in order to get full marks. Specifically, marking points: screenshots for cluster centroids (5%), cluster assignment (4%) and answer for optimal number of clusters (1%).

Marking Criteria for Question 8 (2+ 3 + 3 + 10 = 18%)

a) Up to 2 points can be given based on the number of correct values within the distance matrix.
b) Up to 3 points can be given based on the number of correct cluster assignment
c) Up to 3 points can be given based on the number of correct centroids.
d) Up to 12 points can be given based on the number of correct cluster assignment and centroids.

5. Grading Rubric

These criteria are intended to help you understand how your work will be assessed. They describe different levels of performance of a given criteria.

Criteria are not weighted equally, and the marking process involves academic judgement and interpretation within the marking criteria.

The grades between Pass and Very Good should be considered as different levels of performance within the normal bounds of the module. The Exceptional and Outstanding categories allow for students who, in addition to fulfilling the Excellent requirements, perform at a superior level beyond the normal boundaries of the module and demonstrate intellectual creativity, originality and innovation.

	HONOURS (FHEQ LEVEL 6)
90 +	**Outstanding demonstration of independent scholarly achievement and critical evaluation**. • well-structured assessment that comprehensively addresses the module learning outcomes and specific criteria • critical evaluation is evident through systematic, relevant, and comprehensive coverage of content • skilfully communicated in a style appropriate to the assessment brief • very limited areas for improvement • accurate and consistent use of a recognised referencing system • wide range of current and seminal sources
80 +	**Exceptional demonstration of independent scholarly achievement and critical evaluation**. • well-structured assessment that addresses the learning outcomes and specific criteria for the module • critical evaluation is evident through systematic, relevant and comprehensive coverage of content • skilfully communicated in a style appropriate to the assessment brief • accurate and consistent use of a recognised referencing system • wide range of current and seminal sources
70 +	**Excellent demonstration of independent scholarly achievement and critical evaluation.** • well-structured assessment that addresses the learning outcomes and specific criteria for the module • critical evaluation is evident through systematic and relevant coverage of content • skilfully communicated in a style appropriate to the assessment brief • accurate and predominately consistent use of a recognised referencing system • wide range of appropriate sources
60 +	**Very good demonstration of independent scholarly achievement. and critical evaluation** • well-structured assessment that addresses the learning outcomes and specific criteria for the module • critical evaluation is evident through relevant coverage of content • clearly communicated in a style appropriate to the assessment brief • predominantly consistent and generally accurate use of a recognised referencing system

	HONOURS (FHEQ LEVEL 6)
	• good range of appropriate sources
50 +	**Good demonstration of independent scholarly achievement and critical evaluation**. • fairly well-structured assessment that addresses the learning outcomes and specific criteria for the module • some critical evaluation is evident through coverage of content • good communication in a style appropriate to the assessment brief • predominantly consistent and generally accurate use of a recognised referencing system • a range of appropriate sources
40 +	**Adequate demonstration of independent scholarly achievement and critical evaluation.** • adequately structured assessment that addresses the learning outcomes and specific criteria for the module • some critical evaluation is evident through coverage of content which is also descriptive • communicates in a style appropriate to the assessment brief • attempts to use a recognised referencing system but may have occasional systematic errors • a limited selection of appropriate sources
30 +	**Limited demonstration of independent scholarly achievement and critical evaluation t.** • poorly structured assessment that does not completely address the learning outcomes and specific criteria for the module • work is descriptive in its coverage of the content • Poor communication that does not use a style appropriate to the assessment brief • use of recognised referencing system is systematically inaccurate in a number of places • an insufficient range of appropriate sources
20 +	**Minimal demonstration of independent scholarly achievement and critical evaluation.** • poorly structured assessment that only address a small part of the module learning outcomes and specific criteria for the module • work is descriptive in its coverage of the content, and in places may be inadequate • poor communication that does not use a style appropriate to the assessment brief • use of recognised referencing system is systematically inaccurate throughout the document • an insufficient range of appropriate sources
10 +	• poorly structured assessment that does not address the learning outcomes and specific criteria for the module • coverage of the content is inadequate or incomplete • poor communication that does not use a style appropriate to the assessment brief • recognised referencing system is not used • sources are very limited or absent, or over reliance on one or two sources are very limited or absent, or over reliance on one or two sources
0+	Poorly structured assessment that does not address at all the learning outcomes and specific criteria for the module