Faculty of Science and Technology
Department of Computing & Informatics

Unit Title: Machine Learning (COMP5057)		
Assessment Title: Machine Learning – Heart Disease Classification		
Unit Level: 5	**Assessment Number:** 1 **of** 1	
Credit Value of Unit: 20	**Date Issued:** 27/09/2021	
Unit Leader: Vegard Engen	**Submission Due Date:** 14/01/2022	**Time:** 12:30 PM
Other Marker(s): N/A	**Submission Location:** Turnitin	
Quality Assessor (QA): Avleen Malhi	**Feedback Method:** Brightspace	

This is an individual assignment which carries 100% of the final unit mark.

ASSESSMENT TASK
This ***anonymous*** assignment addresses all Intended Learning Outcomes (ILOs) for this unit (see below).

As context to this assignment, consider the following scenario.

Scenario	Dr Heart from FutureHealth Hospital is keen on improving the detection of heart disease. She has a keen interest in technology and secured some funding to work with a team of Machine Learning consultants and fellow medical experts to create a dataset intended to be used to train a Machine Learning tool to classify whether a patient has got heart disease. Phase 1 of the project is complete, and a dataset has been created. In phase 2, the next steps are to a) perform an exploratory analysis of the data and b) to perform an empirical research experiment to evaluate whether a Machine Learning approach is indeed successful.
Your role	You have a critical role in phase 2 of this project, tasked with doing the exploratory data analysis and conducting the empirical evaluation outlined above.
Key considerations	The final outcomes of your tasks should be written up in a report for Dr Heart and colleagues. While they will not understand all the technical details, they understand data analysis and empirical research well. The report may also be read by technical Data Scientists, who should be able to repeat your experiments in the future to replicate the results if need be.

A key aspect of this assessment is demonstrating the ability to perform a ***critical evaluation*** and ***presenting your findings***. This involves empirical experiments, evaluating the performance of multiple machine learning algorithms and, potentially, data processing techniques, depending on the machine learning workflows you decide to implement/examine.

Based on the scenario above and the dataset you are given (more details below), the main activities are:

a. Conduct an analysis of the dataset and describe its properties.
b. Perform pre-processing of the dataset.
c. Conduct an empirical evaluation of a selection of off-the-shelf machine learning classifiers, e.g., in Scikit-Learn.
d. Present, discuss and reflect on the results from your empirical evaluation.
e. Make conclusions and specific recommendations for potential improvements (to your machine learning workflow).

More details on each of the main activities for this assignment are discussed below, followed by details about the dataset and problem itself.

A - DATA ANALYSIS
The first activity for this assignment is to perform an exploratory data analysis, which is essential because it informs everything you do with Machine Learning in the next steps.

In summary, you should a) give a definition of the problem (for the assignment as a whole), b) give a fact sheet description of the properties of the data set and c) present an analysis of the data (using both statistical and visual techniques).

For the latter (data analysis), you can use the following questions to guide this effort:

- Is the dataset representable for the problem?
- Does the dataset need processing / cleaning?
- Any properties of the dataset that may pose challenges to machine learning classifiers or skew results?
- What insights about the problem can be found in the data?

We will go through details on exploratory data analysis in the first few weeks of the unit.

Have a look at the dataset description further below, and use this GitHub repository as a starting point (contains a Jupyter Notebook that loads the dataset): https://classroom.github.com/a/EBu1oy4x

B – DATA PROCESSING
Not only should the data analysis you do (as per the above) give you and others a real understanding of the data you are working with, but it should directly inform which ***data processing*** you will be doing, which ***classifiers*** you will use (c), as well as particular considerations you need to make when performing the ***critical evaluation of your findings*** (d and e).

For the data processing, you need to consider how you deal with potential issues your exploratory data analysis uncovers. For example, there may be missing values or noise (potentially erroneous data), which you need to deal with. You may also need to scale or transform data for certain machine learning techniques to be able to be applied or perform better, which we will discuss in class (along with other data processing you should consider).

A key to this part of the assignment is deciding and ***specifying*** what data processing methods you will use and ***justify*** your choices. You may, for example, explore how successful different ways of processing the data are as part of your assignment, which is highly encouraged.

C – EMPIRICAL EVALUATION
This part is about running **empirical experiments** to evaluate off-the-shelf machine learning classifiers (and potentially data processing methods), to solve the problem outlined in the scenario above.

For this, you need to choose a set of classifiers and ***justify your choices***, as per the comment above about data processing methods. You also need to give details about how you are configuring the classifiers and how you will evaluate them.

You are expected to evaluate both **effectiveness** and **efficiency**, i.e., considering both classification performance metrics as well as the computational time for training and testing. However, emphasis is on the former (effectiveness).

D – PRESENTING RESULTS
Having run empirical experiments with your chosen machine learning classifiers, you need to present the results in order to answer the main question set out in the scenario above. You need to use appropriate tables and visualisations, and perform a ***critical evaluation*** of the results, discussing pros and cons (advantages and disadvantages).

For this part, it is important to connect the analysis to the real-world problem, considering the implications if the machine learning solution were to be used in a live system.

E – CONCLUSIONS AND FURTHER WORK
Based on your data processing, chosen machine learning classifiers and critical evaluation, you need to draw clear conclusions of whether the problem has been solved successfully or not. You also need to consider improvements to the machine learning workflow as potential future work.

DATASET / PROBLEM

For this assignment, you will work with a dataset for classifying whether patient has got heart disease (or not) from an anonymised dataset. This dataset can be downloaded the GitHub Classroom link provided above.

As it is part of your assignment to describe the dataset, only the following high-level information is given.

The data consist of health records of patients that are classified according to the presence or absence of coronary artery disease; integer value from 0 (absence) to 4 based on the severity of the disease (in the final attribute).

The problem should be solved as a multi-class problem, though you may create a binary version of the dataset for comparison if you wish (combining classes 1-4).

While you are not expected to understand the attributes in terms of their medical meanings they are described as follows:

1. Age
2. Gender
3. Chest pain type
    - 1 = Typical angina
    - 2 = Atypical angina
    - 3 = Non-anginal pain
    - 4 = Asymptomatic
4. Resting blood pressure (in mm Hg on admission to the hospital)
5. Serum cholesterol in mg/dl
6. Fasting blood sugar > 120 mg/dl ? 1 = true, 0 = false
7. Resting electrocardiographic results
    - 0 = normal
    - 1 = having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
    - 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria
8. Maximum heart rate achieved
9. Exercise induced angina ? 1 = yes; 0 = no
10. ST depression induced by exercise relative to rest
11. The slope of the peak exercise ST segment
    - Unsloping
    - Flat
    - Downsloping
12. The number of major vessels (0-3) coloured by fluoroscopy
13. Thalassemia (thal), a blood disorder: 3 = normal; 6 = fixed defect; 7 = reversible defect

Note that this dataset contains missing data, which you will need to handle as part of this assignment.

SUBMISSION FORMAT

You need to submit a technical report:

- Format: PDF, DOC or DOCX
- Word count restriction: 3000 words, which EXCLUDES references, appendices, and tables of content/figures/tables.
- Submission location: Brightspace

The technical report should be written in such a way that it will make sense to anybody who have not read this assignment brief, bearing in mind the scenario provided above. For details about the suggested structure and content of this report, ***please see Appendix A***.

As a technical report, it is expected to have ***numbered headings*** and appropriate ***referencing*** (BU Harvard [1]).

[1] https://libguides.bournemouth.ac.uk/bu-referencing-harvard-style

MARKING CRITERIA

The following criteria will be used to assess the assignment:

Criteria	**Mark**	**ILO(s)**
Professionalism – *structure and quality of the report, including effective use of data visualisation*	15%	3
Data – *description, analysis and insights*	30%	1, 4
Method – *aims of the investigation (optionally hypotheses too), completeness and justification of choices, including data processing, choice of classifiers, parameter tuning, validation approach and evaluation metrics*	25%	1, 2, 3
Results and discussion – *clear narrative, depth of analysis, objective discussion, use of statistical significance, effective use of visualisations and clear connection to the problem (real world implications)*	25%	1, 2, 3, 4
Conclusions and further work – *should include recommendations for improving your machine learning algorithm and the machine learning workflow*	5%	5

To pass this assignment, consider the following as a guideline, though you may excel in different areas in order to bring your mark up.

- Provide a basic description and analysis of the data (e.g., number of instances, types of features, class imbalance, and missing data).
- Specify at least two off-the-shelf classification algorithms chosen for a comparative, empirical, investigation, though the justification of the choice may be limited.
- Specify the chosen validation approach and a selection of standard evaluation metrics discussed in the lectures, though the justification of the choices may be limited. You may not explore ways of handing missing data, feature selection or parameter tuning.
- Results from a comparative study are presented, using both figures and tables, but limited in critical evaluation and discussion.
- Brief conclusions on which algorithm performed best on the dataset, and some lessons learnt.

To obtain high marks for this assignment, consider the following:

- Perform a more in-depth analysis of the data, considering, e.g., noise, outliers and any cleaning that may be required. You also extract and clearly present insights about the problem itself from the dataset.
- Specify one or more secondary objectives, focusing on evaluating different options for data processing (based on your exploratory data analysis), e.g., different methods for dealing class imbalance, missing values, feature selection, feature scaling, handling noise, etc.
- Specify and clearly justify 3+ off-the-shelf classification algorithms chosen for a comparative, empirical, investigation. The justifications will be linked to the data analysis as well as citing literature (e.g., showing that a particular classifier is appropriate and generally performs well).
- Specify and clearly justify the chosen validation approach, methods of handling missing data, feature selection, parameter tuning and evaluation metrics. Again, citing literature to support your justifications.
- Clear and objective presentation of results (figures, tables and narrative), with a critical evaluation of pros and cons, using different evaluation criteria/metrics.

Concise conclusions on algorithm performance (considering statistical significance), lessons learnt and what you would recommend changing or exploring to (potentially) achieve further performance gains.

INTENDED LEARNING OUTCOMES (ILOs)
This unit assesses your ability to:

1. Demonstrate an understanding of basic descriptive statistics, distributions, and their estimation,
2. Demonstrate an in-depth understanding of various machine learning algorithms, and their applicability and limitations for problems in a real-world context,
3. Demonstrate the ability to confidently use essential software products relevant to machine learning and data visualisation,
4. Design and evaluate complex machine learning workflows, and
5. Rigorously analyse performance of machine learning models and recommend actions to improve them,

QUESTIONS ABOUT THE BRIEF
This assignment will be discussed in class, where students are encouraged to ask questions for clarification. Some common questions and answers are included in ***Appendix B***. Questions will also be answered via Microsoft Teams (details on Brightspace).

Unit Leader Signature Vegard Engen

Help and Support

Undergraduate Coursework Assessments

If a piece of coursework is not submitted by the required deadline, the following will apply:

1. If coursework is submitted within 72 hours after the deadline, the maximum mark that can be awarded is 40%. If the assessment achieves a pass mark and subject to the overall performance of the unit and the student's profile for the level, it will be accepted by the Assessment Board as the reassessment piece. The unit will count towards the reassessment allowance for the level; This ruling will apply to written coursework and artefacts only; This ruling will apply to the first attempt only (including any subsequent attempt taken as a first attempt due to exceptional circumstances).
2. If a first attempt coursework is submitted more than 72 hours after the deadline, a mark of zero (0%) will be awarded.
3. Failure to submit/complete any other types of coursework (which includes resubmission coursework without exceptional circumstances) by the required deadline will result in a mark of zero (0%) being awarded.

The Standard Assessment Regulations can be found on **Brightspace or via** https://www1.bournemouth.ac.uk/students/help-advice/important-information (under Assessment).

Exceptional Circumstances

If you have any valid **exceptional circumstances** which mean that you cannot meet an assignment submission deadline and you wish to request an extension, you will need to complete and submit the online Exceptional Circumstances Form together with appropriate supporting evidence (e.g. GP note) normally **before the coursework deadline**. Further details on the procedure and links to the exceptional circumstances forms can be found on **Brightspace or via** https://www1.bournemouth.ac.uk/students/help-advice/looking-support/exceptional-circumstances. Please make sure that you read these documents carefully before submitting anything for consideration. For further guidance on exceptional circumstances please contact your Programme Leader.

Referencing

You must acknowledge your source every time you refer to others' work, using the **BU Harvard Referencing** system (Author Date Method). Failure to do so amounts to plagiarism which is against University regulations. Please refer to https://libguides.bournemouth.ac.uk/bu-referencing-harvard-style for the University's guide to citation in the Harvard style. Also be aware of Self-plagiarism, this primarily occurs when a student submits a piece of work to fulfill the assessment requirement for a particular unit and all or part of the content has been previously submitted by that student for formal assessment on the same/a different unit. Further information on academic offences can be found on **Brightspace** and from https://www1.bournemouth.ac.uk/discover/library/using-library/how-guides/how-avoid-academic-offences

Additional Learning Support

Students with **Additional Learning Needs** may contact the Additional Learning Support Team. Details can be found here: https://www1.bournemouth.ac.uk/als

Primary Research (Undergraduate Levels)

You should not be conducting any primary research (i.e. carrying out an investigation to acquire data first-hand, for example, where it involves approaching participants to ask questions or to participate in surveys, questionnaires, interviews, observations, focus groups, etc.) unless otherwise specified in the brief. However, if there is a genuine requirement to collect primary research data you will require ethical approval before doing so. In the first instance, please discuss with the Unit Leader. The collection of primary data without appropriate ethical approval is a serious breach of Bournemouth University's Research Ethics Code of Practice and will be treated as Research Misconduct.

IT Support

If you have any problems submitting your assessment please contact the IT Service Desk - +44 (0)1202 965515 - immediately and before the deadline.

Disclaimer

The information provided in this assignment brief is correct at time of publication. In the unlikely event that any changes are deemed necessary, they will be communicated clearly via e-mail and Brightspace and a new version of this assignment brief will be circulated.

APPENDIX A – TECHNICAL REPORT STRUCTURE

Expected structure of the report:

1. Front matter
    a. Title page (title & date[1])
    b. Table of content
    c. Table of figures
    d. Table of tables
    e. List of definitions and abbreviations
2. Exploratory data analysis
3. Method (see details below)
4. Results and discussion
5. Conclusions and future work (lessons learnt / recommendations for improvements)
6. References
7. Appendices

For the method, you need to

a) Define the aims and objectives of the investigation (use SMART[2] objectives, with well-defined success criteria)
b) Specify and justify data processing methods you will use and evaluate (linked to your data analysis)
c) Specify and justify the classification algorithms you will use and evaluate
    - The description(s) should be brief, citing one or more sources for a complete description
d) Specify and justify the chosen validation method (e.g., cross validation or holdout validation)
e) Specify and justify other evaluation aspects, such as the configuration and potential parameter tuning of the classification algorithms (to ensure a fair comparison), and the metrics used to evaluate performance

[1] As this is an anonymous assessment, please do not include your name or student ID
[2] https://en.wikipedia.org/wiki/SMART_criteria

APPENDIX B – FREQUENTLY ASKED QUESTIONS

Q: Can I exceed the word count? Is 10% extra OK?
A: No. Instead, you need to work on writing succinctly and using appendices effectively.

Q: Can I put my data description in a table and use show it as an image because it is taking up a chunk of my word count?
A: No, sorry, taking images of text is not an acceptable approach here. However, you can use an appendix for **non-essential** details such as describing the features of the dataset. Though, anything that is marked directly should be in the main body of the text.

Q: What do you mean by empirical experiment?
A: Empirical means you get empirical evidence from direct observation of an experiment. In practical terms here, it means you run machine learning algorithms against the dataset provided and get quantitative data from.

Q: What do you mean by an off-the-shelf classifier?
A: This is a classifier that's already been implemented by somebody else and is readily available for you to use, e.g., from Scikit-Learn.