Faculty of Computing, Engineering and Science (CES)

Assessment Cover Sheets

University of
South Wales
Prifysgol
De Cymru

Faculty of Computing, Engineering and Science (CES)

Module Title	**Applied Machine Learning and Deep Learning**
Module Code	**MS4S16**
Module Leader/Tutor	**Sam Jobbins**
Assessment Type	**Practical Coursework 1 (Asynch)**
Assessment Title	**Assessment 1**
Word Count/Duration/Equivalent	**N/A**
Submission Date	**6th February 2025**
Return Date	**20th March 2025**
Weighting	**50%**

Faculty of Computing, Engineering and Science (CES)

Assessment Cover Sheets

Assessment Description

You are required to apply the data pre-processing techniques, the feature engineering techniques and the machine learning models seen in this module to a given dataset.

You are also required to produce a formal report summarising your findings for the tasks detailed above. You can apply any and as many machine learning algorithms and/or pre-processing steps as you wish, as long as the techniques and steps you are taking are appropriate for the task.

Your report should be written clearly and legibly and your code should follow best practices.

Further information and a full description are provided following the Assessment Cover Sheets.

Guidance on Format of Assessment

The assessment is broken down into four sections: Sections 1, 2 and 3 are individual components in which you will be asked to perform particular tasks; Section 4 refers to the whole report, and includes marks on presentation, good coding practices and clear and concise explanations. Further information and a full description are provided following the Assessment Cover Sheets.

You must complete this work using Python and the report should be created in Jupyter Notebooks, using appropriate text/markdown cells with explanations, code snippets, comments and visualisations.

Once completed, you should then export your Jupyter Notebook file as a .pdf or an .html file, which you will then upload to Turnitin *via* your Blackboard module.

Learning Outcomes Assessed

- To understand the concepts of machine learning and deep learning for Data Science, and compare and test a range of techniques;
- To classify features of data sources, analysing and interpreting the outputs of machine learning and deep learning techniques in the context of practical situations in the area of Data Science.

These learning outcomes, and more information about them, is specified in the validated module descriptor https://curriculum.southwales.ac.uk/Module/Details?moduleId=MOD012335.

Marking Criteria/Rubric

The assessment is broken into four sections. Marks will be broken down as follows:

- Section 1 is worth **35 Marks;**
- Section 2 is worth **25 Marks;**
- Section 3 is worth **25 Marks;**
- Section 4 is worth **15 Marks.**

The total for this assessment is therefore **100 Marks**, which represents 50% of the module. More details on each section, and an overall marking rubric giving information about how these marks are awarded, is given below.

Note: All grades are provisional until they are ratified by the exam board.

Supporting your Success

Here are some resources to support you to be successful:

Assignment Tips and Wellbeing Treats

Using Feedback to Improve

Left it to the Last Minute

Submission Details

Once you have completed the assessment, please submit a **.pdf or .html version of your report** *via* Turnitin, which can be found on your Blackboard module under the *Assessments* section**.**

Please note that Jupyter Notebook files (.ipynb), Word Document (.doc, .docx) and other file types will **NOT** be accepted.

Please ensure that your final document's filename contains your name and student number.

It is your responsibility to ensure that your submission has been uploaded to the correct location and in the correct filetype. Failure to do so may lead to a deduction of marks.

What happens next?

Your marked assessment should be available 20 working days after submission. However, please be advised that this may be subject to change in the event of Bank Holidays, University Closure or staff sickness. If there is something about the feedback you have been given that you are unclear about, please see your module tutor.

Feedback Method

Feedback will be provided as a percentage score, accompanied by comments on your individual Turnitin assessment submission.

Late Submission

You are allowed a further five working days from the submission deadline in which to submit, however, your mark will be capped at the pass mark (usually 40%). If you are a student who has an Individual Support Plan (ISP), your mark will not be capped.

What happens if you don't submit?

If you do not submit your assessment, or effectively claim Extenuating Circumstances you will receive 0%. This may affect your chances of successfully passing the module. It may also be used as an indicator that you are not engaging on your course, and you may be referred to the Lack of Engagement Process.

In-Year Retrieval in the Event of Failure

Assessments on this module are eligible for In-Year Retrieval (IYR), meaning that if you fail the assessment, you will have another chance to pass, with the mark being capped at the pass mark of 40%. Please ask your module tutor or your course leader for more information.

Retrieval in the Event of Failure

If the assessment is failed and In-Year Retrieval (IYR) is not successful, not taken or not offered, you may need to resit this assessment during the resit period.

Extenuating Circumstances

If you are experiencing circumstances that are affecting your performance in assessments, you may be able to claim Extenuating Circumstances. You may need evidence of your circumstances to make a claim. Further information can be found on the UniLife Extenuating Circumstances pages. If you need support to make a claim, Advice Zone Online has a helpful FAQ about Extenuating Circumstances. You can 'ask a question' which will be sent to Advice Zone staff. You can also Contact the Advice Zone directly.

Referencing, Plagiarism and Good Academic Practice

Please remember to reference any external sources you may use for your information using an appropriate referencing style. Failure to adequately reference work from other sources or obvious incidents of copying between students will be considered as plagiarism and may lead to a loss of marks.

Please also make sure you are aware of the University's policies on fair use of Artificial Intelligence (AI). You are permitted to use AI to generate ideas and explore concepts (providing that you reference the tool and explicitly provide the prompts you have used in your references section), however directly copying and pasting output from AI sources such as ChatGPT is considered plagiarism. Please refer to the University's policies on fair use of AI for further information.

For further information about Good Academic Practice, visit: https://advice.southwales.ac.uk/a2z/referencing-plagiarism-and-good-academic-practice

Feeling overwhelmed?

USW's Wellbeing Service offer free advice and support to all USW students.

You will also find links to self-help and can access one of their specialist services:

- Mental Health Service
- Counselling
- Health Service
- Disability

For other help and support check the USW support services page.

Learning Support Resources

The University's Study Skills page has a large number of Learning Support Resources available: https://studyskills.southwales.ac.uk

Your Assessment Queries

If you have any questions about the instructions or the submission of this report, please contact sam.jobbins@southwales.ac.uk.

MS4S16 Assessment 1

Submission Deadline: Thursday, 6th February 2025 11:59am (11:59)

Contribution to the Module: 50%

If you have any questions about the instructions or the submission of this report, please contact sam.jobbins@southwales.ac.uk

Instructions

You are required to apply the data pre-processing techniques, the feature engineering techniques and the machine learning models seen in this module to a given dataset.

You are also required to produce a report summarising your findings for the tasks detailed above. You can apply any and as many machine learning algorithms and/or pre-processing steps as you wish, as long as the techniques and steps you are taking are appropriate for the task. Your report should be written clearly and legibly and your code should follow best practices.

You MUST use the version of the dataset supplied. Failure to do so will result in a significant loss of marks.

You must complete this work using Python and the report should be created in Jupyter Notebooks, using appropriate text/markdown cells with explanations, code snippets, comments and visualisations.

Submission

Once you have completed the assessment, please submit a **.pdf or .html version of your report** *via* Turnitin, which can be found on your Blackboard module under the *Assessments* section**.**

Please note that Jupyter Notebook files (.ipynb), Word Document (.doc, .docx) and other file types will **NOT** be accepted.

Please ensure that your final document's filename contains your name and student number.

It is your responsibility to ensure that your submission has been uploaded to the correct location and in the correct filetype. Failure to do so may lead to a deduction of marks. If in doubt, ask!

MS4S16 – Applied Machine Learning and Deep Learning (2024-2025)

Assessment 1

Plagiarism and Academic Misconduct

Please remember to reference any external sources you may use for your information using an appropriate referencing style. Failure to adequately reference work from other sources or obvious incidents of copying between students will be considered as plagiarism and may lead to a loss of marks and/or a referral to the Academic Misconduct team.

Please also make sure you are aware of the University's policies on fair use of Artificial Intelligence (AI). You are permitted to use AI to generate ideas and explore concepts (providing that you reference the tool and explicitly provide the prompts you have used in your references section), however directly copying and pasting output from AI sources such as ChatGPT is considered plagiarism. Please refer to the University's policies on fair use of AI for further information.

Information on the Dataset

The rapid advancement of advanced machine learning (ML) techniques has transformed how data is analysed and interpreted across various scientific disciplines. One such area is computational toxicology, a cross-disciplinary field at the interface of computer science, chemistry, pharmacology and medicine.

Computational toxicology uses data-driven and informatics-based techniques to studies of how chemical, physical and biological agents affect living organisms and the environment. An important measure of the toxicity of a chemical agent is its median lethal dose (LD_{50}), which represents the dose of a substance that causes death in 50% of a test population; a lower LD_{50} means that the substance has a higher acute toxicity. This serves as a key indicator for assessing the safety of pharmaceutical compounds and environmental chemicals.

You are provided with a dataset entitled '`MS4S16 – Assessment 1 – LD50 Dataset.csv`', which is saved in Comma-Separated (CSV) format. This file contains a modified version of a dataset containing LD_{50} values of 7,413 compounds.

The dataset is derived from the paper of Wu, Kedi and Guo-Wei Wei, who used deep learning to assess toxicity of chemical compounds. The original dataset only contained the SMILES codes (a cheminformatics notation used to identify compounds) for each chemical, and so the dataset has been modified by Sam to include many other features using the *rdkit* and *pubchempy* Python libraries. **Because of this, make sure you use the version of the dataset provided on Blackboard, and not a different version of the dataset from the internet.**

The data contains 7,413 rows (each a different chemical compound) and 28 columns, including IUPAC name (the name of the chemical), SMILES identifier twenty-five descriptors and the value of LD_{50} itself). Some information on the columns can be found below:

Feature Name	Feature Description
Name	IUPAC name of the chemical compound
SMILES	SMILES identifier – a unique string encoding the structure of each chemical compound
LD50	The median lethal dose (LD_{50}) value of the compound. In your supervised analysis, this will be your target variable.

BertzCT	Descriptor from J. Am. Chem. Soc. 103:3599-601 (1981)
Chi2v	Descriptor from Rev. Comput. Chem. 2:367-422 (1991)
Chi3v	Descriptor from Rev. Comput. Chem. 2:367-422 (1991)
Chi4n	Descriptor from Rev. Comput. Chem. 2:367-422 (1991)
Chi4v	Descriptor from Rev. Comput. Chem. 2:367-422 (1991)
HeavyAtomCount	Number of heavy atoms in the molecule
HeavyAtomMolWt	Molecular weight of the heavy atoms in the molecule
Kappa3	Descriptor from Rev. Comput. Chem. 2:367-422 (1991)
MaxPartialCharge	Maximum Partial Charge
MinAbsPartialCharge	Minimum Absolute Partial Charge
MinEStateIndex	Descriptor from JCICS 31:76-81 (1991)
MinPartialCharge	Minimum Partial Charge
MolLogP	Partition coefficient from Wildman and Crippen JCICS 39:868-73 (1999)
MolWt	Molecular weight of the molecule
NumHAcceptors	Number of hydrogen bond acceptors in the chemical compound
NumHDonors	Number of hydrogen bond donors in the chemical compound
NumHeteroatoms	Number of heteroatoms (atoms other than carbon or hydrogen)
RingCount	Number of rings in the chemical compound
SMR_VSA10	MOE-type descriptor using MR contributions and surface area contributions
SlogP_VSA12	MOE-type descriptor using LogP contributions and surface area contributions
SlogP_VSA5	MOE-type descriptor using LogP contributions and surface area contributions
VSA_EState4	MOE-type descriptors using EState indices and surface area contributions
VSA_EState9	MOE-type descriptors using EState indices and surface area contributions
qed	Quantiative Estimation of Drug-Likeness: Nature Chemistry, 4, 90-98 (2012)

Note that understanding the detailed meaning of the columns is not strictly necessary and you need only be able to use them in your analysis – this is a data science assessment, not a chemistry or pharmacology test! However, if you are interested, further information on molecular descriptors used in the *rdkit* library can be found here: https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors

For further information on the original dataset and paper, you can find the publication here: Wu, Kedi, and Guo-Wei Wei. "Quantitative toxicity prediction using topology based multitask deep neural networks." Journal of Chemical Information and Modeling, 58, no. 2, 520-531 (2018) https://doi.org/10.1021/acs.jcim.7b00558.

Aims of the Assessment

To solve this assessment, you may wish to use any of the techniques that we have learned in the first half of the module, as well as using other concepts and ideas you find through independent study.

You are, of course, welcome to look for help and inspiration online, but refrain from copying code verbatim from other sources and always remember to reference your sources – failure to do so is plagiarism!

The marks of the coursework constitute 50% of the mark of this module

The marks for this work can be broken down into four sections, each worth a different percentage of the overall grade. The percentage for each section is given below:

---

1. **Pre-process the dataset and perform an Exploratory Data Analysis (EDA) of the data. (35%)**

This should include:

- Splitting the dataset into a training set and a test set;
- Taking care of any missing, duplicated or outlier values;
- Transforming data, where appropriate to do so;
- Appropriately treating or encoding text and categorical features;
- Performing feature engineering techniques such as feature creation, extraction and selection;
- Looking at the relationships between variables in your training set;
- Producing appropriate and informative plots and tables for an exploratory analysis, and justifying the use of particular plots;

---

2. **Utilising features and attributes derived from the pre-processing and EDA stage, conduct an unsupervised machine learning analysis with the aim of gaining further insights into the data via clustering or dimensionality reduction.**
   (25%)

To do this, you may consider:

- Clustering using different appropriate algorithms, e.g. K-means, hierarchical, DBScan;
- Performing a dimensionality reduction to see if a smaller number of features can adequately explain the observations;
- Evaluating the utility of the different algorithms with appropriate metrics;
- Drawing some informative visualisations and plots, if appropriate to do so;
- Critically assessing your statistical assumptions and inferences.

---

3. **Utilising features and attributes from your pre-processing, EDA and unsupervised work, conduct a supervised machine learning analysis with the aim of predicting the median lethal dose (LD_{50}) variable using regression techniques.**
   (25%)

To do this, you may think of:

- Creating a series of regression models to predict the numerical features;
- Evaluating your supervised models with metrics of your choice;
- Optimising, tuning and validating your models with parameter tuning and cross-validation;
- Drawing some informative visualisations and plots, if appropriate to do so;
- Critically assessing your statistical assumptions and inferences.

---

4. **Provide well written paragraphs, annotations, tables and plots in your Jupyter Notebooks explaining in detail the workflow, the results and the reasons for the choices you have made.**
   (15%)

- At the end of each part of the analysis, summarise and reflect on the task at hand. Use the opportunity to present and support your decisions with informative tables, graphs, and explanations;
- Outline the structure of the report in a logical and flowing fashion, as you might expect of a workflow in the context of a professional presentation or academic publication;
- Conclude your notebook with a final conclusions, evaluations and recommendations section, drawing all of the analysis together and highlighting any limitations of the work done;

- Discuss briefly any ethical implications you can think of arising from this kind of work.

Closing Remarks

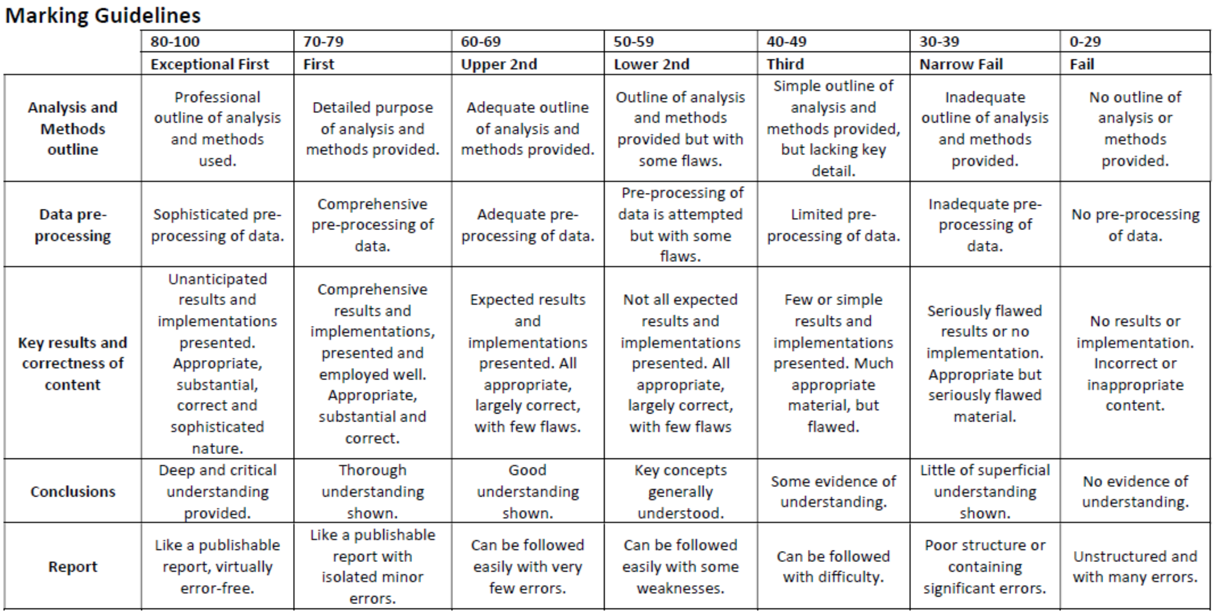
This is a challenging piece of work that gives you an opportunity to showcase everything you have learned in the last few weeks: Don't panic! There is nothing here that you cannot do with the expertise you will have developed over this course.

If you are not sure about something… ask me or one of the course team, use your lecture notes, the tutorial code books, the asynchronous materials, or the recommended reading list, to try and find out more.

There will also likely be lots of useful hints and tips online, particularly at websites like Towards Data Science, Stack Overflow and Kaggle – but as I said before, please do not just copy other peoples' solutions directly from the internet, as this is plagiarism!

If you have any further questions, drop me an email at sam.jobbins@southwales.ac.uk or come and see me in my office (TR J 4 18)

GOOD LUCK – POB LWC!



Marking Guidelines

	80-100	70-79	60-69	50-59	40-49	30-39	0-29
	Exceptional First	**First**	**Upper 2nd**	**Lower 2nd**	**Third**	**Narrow Fail**	**Fail**
Analysis and Methods outline	Professional outline of analysis and methods used.	Detailed purpose of analysis and methods provided.	Adequate outline of analysis and methods provided.	Outline of analysis and methods provided but with some flaws.	Simple outline of analysis and methods provided, but lacking key detail.	Inadequate outline of analysis and methods provided.	No outline of analysis or methods provided.
Data pre-processing	Sophisticated pre-processing of data.	Comprehensive pre-processing of data.	Adequate pre-processing of data.	Pre-processing of data is attempted but with some flaws.	Limited pre-processing of data.	Inadequate pre-processing of data.	No pre-processing of data.
Key results and correctness of content	Unanticipated results and implementations presented. Appropriate, substantial, correct and sophisticated nature.	Comprehensive results and implementations, presented and employed well. Appropriate, substantial and correct.	Expected results and implementations presented. All appropriate, largely correct, with few flaws.	Not all expected results and implementations presented. All appropriate, largely correct, with few flaws	Few or simple results and implementations presented. Much appropriate material, but flawed.	Seriously flawed results or no implementation. Appropriate but seriously flawed material.	No results or implementation. Incorrect or inappropriate content.
Conclusions	Deep and critical understanding provided.	Thorough understanding shown.	Good understanding shown.	Key concepts generally understood.	Some evidence of understanding.	Little of superficial understanding shown.	No evidence of understanding.
Report	Like a publishable report, virtually error-free.	Like a publishable report with isolated minor errors.	Can be followed easily with very few errors.	Can be followed easily with some weaknesses.	Can be followed with difficulty.	Poor structure or containing significant errors.	Unstructured and with many errors.