COMP3222 Machine Learning Technologies (UoSM) 2025/2026
Coursework I

Module Code:	COMP3222		
Module Title:	Machine Learning Technologies		
Module Leader:	Asst.Prof. Dr Hasmath Farhana Thariq Ahmed		
Assessment Type:	Individual	**Weightage:**	50%
Submission Due Date:	**08-Dec-2025, 12 PM (MYT)**		
Method of Submission:	**Blackboard**		
GenAI **Tier Specification**	**Tier 1: No GenAI use [applies All - Entire work]**		

This assessment relates to the following Module/Course Learning Outcomes (CLO):

Numeracy Skills	CLO2	Adopt a systematic approach to data analysis within learning algorithms, aiming to uncover new patterns or concepts, and subsequently analyse their performance using a scientific computing environment. (A3, PLO7)

COMP3222 Machine Learning Technologies (UoSM) 2025/2026
Machine Learning Technologies Coursework

This Coursework is worth **50%** of the total marks for this module. The deadline is **Monday, 08-Dec-2025, 5 PM (MYT), hand in' via the Blackboard** site. Depending on your prior knowledge, the work will take 35-40 hours approximately.

Coursework Overview:

This coursework **requires a basic understanding of classical machine learning algorithms** and the ability to differentiate between a regression and classification problem to choose the appropriate model for the given problem. You are expected to understand how to select appropriate assessment metrics to demonstrate the performance of the chosen algorithm. Additionally, your knowledge of visualisation tools will aid in providing a comparative performance analysis.

Coursework Requirements

- You are expected to present adequate information about your work in a PDF report and include anything else you think would be useful in this context in the appendix section. For example, every detail starting from the dataset, including the preprocessing, assessment metrics and performance analysis of your implemented learning model and its parameters, is to be clearly recorded with illustrations wherever required, with supporting justifications pertaining to the page limit mentioned.
- Make sure you record URLs/details of the sources you used and cite them. For example, all sources have been properly referenced and cited using the citation style* (Ahmed, Ahmad et al. 2020).
- The PDF report must be submitted only through Blackboard.

Notebook

- You can certainly use ***Jupyter notebook*** / ***Google Colab notebook*** as your logbook for recording your coursework progress, answers, thoughts, reflections, and any questions that arise during the coursework.
- Save, organise and convert your notebook (.ipynb) file into PDF version in any shared drive (One note/ Google drive) for easy access and reference.
- You are required to add the accessible link of the Jupyter notebook /Google Colab notebook in the appropriate APPENDIX section of the final report.

Although the academics wish to see your notebook (PDF version), its primary function is for you. Completing these tasks will help you grow as an independent learner. If you think about this work, you will inevitably have questions not addressed here. You should also take notice of and record the responses to those questions.

Academic Integrity

"In this entire coursework, students are **prohibited** from **using any GenAI tools** for their **assessed work**. This includes entering any part of the assignment or your assessed work to GenAI, whether by pasting/typing, uploading files, or describing content directly or through plugins. Basic tools that assist spelling and grammar, translation and calculation without generating new content or ideas, can be used unless specified otherwise by the assessment setter. GenAI may be used to explain lecture slides and notes to enhance understanding of a relevant topic areas. Students are not required to complete a GenAI Declaration Form."

For more information on GenAI refer here

The focus of this coursework is learning. While collaboration with peers is encouraged, this is an individual task. Sharing or reproducing others' ideas as your own is strictly prohibited. It's acceptable and encouraged for you to refer to certain web stuff in your notebook. However, in such cases, you must credit the original author's information or the source you referred to. If you cite/acknowledge any source that assists you in obtaining relevant knowledge for completing this coursework, it is highly appreciated, and no marks will be deducted for the same.

Help and Assistance
During lab hours, academics will be on hand to assist you. After the lab, you will need to continue on your own schedule. You are urged to speak with your colleagues about this work and share information and thoughts with them. Nevertheless, you need to record your own findings in the report.

Marking Guidelines & Feedback
This coursework is worth 50% of the total marks for this module. It will be assessed using the attached assessment rubrics. Marks with written feedback will be available within 3 weeks of the coursework due date.

Coursework

The coursework is briefed in the following sections:

1. **Problem Statement:**
   You are provided with a dataset named movie metadata.csv (accessible in the BB on the same day as the handout) that contains several movie-related variables, including the IMDB rating. Your task is to identify a learning model capable of categorising movies into five categories using the properties of the provided information. The classification should adhere to the following criteria:
   IMDB score 1-2: Categorized as "Poor" movies.
   IMDB score 2-4: Categorized as "Below Average" movies.
   IMDB score 4-6: Categorized as "Average" movies.
   IMDB score 6-8: Categorized as "Good" movies.
   IMDB score 8-10: Categorized as "Excellent" movies.

2. **Dataset Description:**
   The dataset contains 28 variables for 5,043 movies, spanning over 100 years in 66 countries. There are 2,399 unique director names and thousands of actors/actresses. The "imdb_score" is the response variable, while the other 27 variables are potential predictors.

   movie_title: Title of the Movie.
   duration: Duration of the movie in minutes.
   director_name: Name of the Director of the Movie.
   director_facebook_likes: Number of likes on the Director's Facebook Page.
   actor_1_name: Primary actor starring in the movie.
   actor_1_facebook_likes: Number of likes on Actor 1's Facebook Page.
   actor_2_name: Other actor starring in the movie.
   actor_2_facebook_likes: Number of likes on Actor 2's Facebook Page.
   actor_3_name: Other actor starring in the movie.

actor_3_facebook_likes: Number of likes on Actor 3's Facebook Page.
num_user_for_reviews: Number of users who gave a review.
num_critic_for_reviews: Number of critical reviews on IMDb.
num_voted_users: Number of people who voted for the movie.
cast_total_facebook_likes: Total number of Facebook likes for the entire cast of the movie.
movie_facebook_likes: Number of Facebook likes on the movie's page.
plot_keywords: Keywords describing the movie plot.
facenumber_in_poster: Number of actors featured on the movie poster.
color: Film colorization (e.g., 'Black and White' or 'Color').
genres: Film categorization (e.g., 'Animation', 'Comedy', 'Romance', 'Horror', 'Sci-Fi', 'Action', 'Family').
title_year: The year in which the movie was released (ranging from 1916 to 2016).
language: Language of the movie (e.g., English, Arabic, Chinese, French, German, Danish, Italian, Japanese, etc.).
country: Country where the movie was produced.
content_rating: Content rating of the movie.
aspect_ratio: Aspect ratio in which the movie was made.
movie_imdb_link: IMDb link of the movie.
gross: Gross earnings of the movie in Dollars.
budget: Budget of the movie in Dollars.
imdb_score: IMDb Score of the movie on IMDb.

3. Coursework Requirements:

The coursework requirements, that must be included in the report are explained in the following subsections i through x.

i. ***Data Cleaning/Preparation:*** You are required to **identify** and apply appropriate data quality principles while loading and cleaning/preprocessing the dataset, including handling missing values, encoding categorical variables, and scaling numerical features. Pay particular attention to data quality, making necessary cleaning and transformation decisions.

ii. ***Feature Selection:*** **Perform** feature selection or dimensionality reduction to choose the most important features influencing IMDb scores. Provide a clear rationale for feature selection, explaining why certain features are chosen and others excluded.

iii. ***Model Selection:*** **Choose** and implement **at least THREE (3) distinct machine learning algorithms** as a solution for the given problem statement as mentioned in Section.1. These models should represent a variety of approaches to movie rating prediction and categorization for the given problem.

iv. ***Model Training:*** Train machine learning models using a portion of the dataset. You should **perform** a suitable train-test split, reserving a portion of the data for model evaluation. **Explain** the reasoning behind the chosen split ratio (e.g., 70/30 or 80/20) and provide a detailed description of the training process for your chosen models.

v. ***Cross-Validation:*** **Perform** cross-validation techniques to assess the models' performance more robustly. Emphasize the importance of k-fold cross-validation to ensure a fair evaluation and mention any strategies used for handling class imbalance, if applicable.

vi. ***Model Evaluation:*** **Describe** the performance of each model using appropriate metrics such as accuracy, precision, recall, and F1-score (all/whichever applicable). Use both the train-test split and cross-validation to thoroughly assess the models. **Discuss** the choice of evaluation metrics and how they relate to the given problem objectives.

vii. ***Model Comparison:*** **Differentiate** the performance of the **THREE (3)** models as you chosen and identify the one that performs comparatively better in terms of IMDb score prediction and movie categorization. Highlight the trade-offs and strengths of each model based on your performance analysis.

viii. ***Model Hyperparameter Tuning:*** **Perform** hyperparameter tuning for the selected model to optimize its performance further. Clearly document the hyperparameters tested and the rationale behind their selection.

ix. ***Model Interpretation:*** Interpret the trained model to understand influential features and predict IMDb scores. **Discuss** how the model's predictions align with the problem statement/objectives and provide insights into the importance of different features.

x. ***Categorization:*** **Categorize** movies into specified categories based on IMDb scores: "Poor," "Below Average," "Average," "Good," and "Excellent."

The above subsection is provided for guiding you with the workflow, however it is your responsibility to execute the steps in the right order or perform the process as appropriate based on your understanding. If you would like to perform any other steps apart from those mentioned in subsections (i. through x.), it is mandatory for you to provide a short reasoning **(not exceeding 200 words in your report)** of doing so.

4. *Report:*

Write a comprehensive **report** in the form of **PDF document**, summarizing the project, including data preprocessing, model selection, evaluation, model comparison, and conclusions (appropriately as mentioned in Section.3).

Stress the importance of transparent and informative reporting to convey your findings effectively.

The following sections are required to be in the report,
a. **Introduction (Maximum 1 Page):** You may consider outlining briefly the problem statement, dataset, and objective of the work performed (whichsoever as applicable from Section 3).

b. **Methodology (Maximum 4 Pages):** You may consider explaining your workflow, that includes - data cleaning/preparation/feature selection/learning models/parameter tuning/all other details regarding the model evaluation/validation (whichsoever as applicable from Section 3).

c. **Results and Interpretation (Maximum 2 Pages):** You may consider explaining model comparison/interpretation/categorization/evaluation with visual representations (tables/graphs/figs) as appropriate. You are required to provide a brief analysis on pros and limitations of each algorithm (whichsoever as applicable from Section 3).

d. **Summary (Maximum 1 Page):** Summarize your findings and emphasise the best performing algorithms (based on the results/interpretation using appropriate metrics) (whichsoever as applicable from Section 3).

e. **References:** Cite all sources used for this report, preferably text citations*.

f. **Appendix (Not exceeding 4 Pages)**
The Appendix must consist of the following

A1. Jupyter Notebook/Google Colab Link
- Convert the jupyter notebook to PDF and present only the URL to the PDF document.

A2. Any other contents (You can include any screenshots or other materials that you feel would be helpful to communicate your report effectively) and refer the same wherever required in your report.

You are required to include details of your coursework requirement as in Section 3, as applicable under appropriate section in the final report with appropriate titles/subtitles. Your final report must not include any screenshots of the **Jupyter notebook/Google Colab codes**. The report is required to include only meaningful **tables/graphs/figures** that effectively communicates your understanding, or your interpretation based on your analysis. Report documentation must be aligned and content to be justified, with **arial font 12, one line spacing.**

*The citation style to be followed in the report
"Device free human gesture recognition using Wi-Fi CSI: A survey." Engineering Applications of Artificial Intelligence **87**: 103281.

APPENDIX - A

Marking Guidelines

Grading Criteria	CLO	Wtg	Excellent 5	Above average 4	Average 3	Fair 2	Low 1	0	Mark (≈Wtg * Score)
Data Preprocessing and feature selection	CLO2	2	Exceptional data preparation with meticulous handling of missing values (if any), encoding of categorical variables, and scaling of numerical features. Demonstrates a deep understanding of data quality. Exceptional feature selection with a clear rationale for each feature chosen and excluded, showcasing a profound understanding of feature importance.	Proficient data preparation with effective handling of missing values (if any) & thoughtful feature selection with explanations highlighting the importance of chosen features.	Competent data preparation but with minor issues such as incomplete handling of missing values (if any). Competent feature selection with brief explanations or minor omissions in rationale.	Data preparation is attempted but contains significant errors, or several important data quality issues remain unaddressed. Feature selection is attempted but lacks a clear rationale or contains significant omissions.	Data preparation is rudimentary, and many data quality issues are not resolved. Feature selection is rudimentary, and the rationale is unclear or absent.	No data preparation or feature selection attempted or no understanding of the importance of data quality and feature selection.	/10
Model Selection (At least three ML algorithms)	CLO2	4	Exceptional choice of algorithms (3 or more) with a profound understanding of their applicability with no errors and better performance.	Thoughtful choice of algorithms with explanations highlighting their suitability with appropriate selection of parameters with 3 different ML algorithms	Competent implementation of two types of supervised algorithms, with some room for improvement.	Solid implementation of supervised algorithms with few errors / Implemented only one algorithm with no errors	Implementation of one or more algorithms is attempted but contains significant errors or issues.	No implementation of supervised machine learning algorithms	/20
Model training, cross validation, evaluation and categorization	CLO2	4	Exceptional model training with a suitable train-test split ratio (e.g., 70/30 or 80/20) explained in detail, and a well-	Proficient model training with a clear rationale for the chosen split ratio and a well-	Competent model training with minor issues in the explanation of the split ratio and	Model training is attempted but contains significant errors, or the split ratio and training	Rudimentary model training with major issues or a lack of clarity in the split ratio and training process.	No model training, cross-validation, model evaluation, or categorization attempted, or a complete lack of	/20

			documented training process. Application of k-fold cross-validation techniques to ensure a fair evaluation, with a strong emphasis on model robustness. Thorough model evaluation using appropriate metrics such as accuracy, precision, recall, and F1-score. Successful categorization of movies into specified categories with detailed reasoning and metrics for each category	described training process. Effective application of k-fold cross-validation for robust evaluation. Proficient model evaluation using appropriate metrics. Successful categorization of movies with clear reasoning and metrics	training process. Competent application of k-fold cross-validation. Competent model evaluation with minor issues. Competent categorization of movies with brief reasoning and metrics.	process lack clarity. Attempted application of k-fold cross-validation but with significant issues. Model evaluation is rudimentary or lacks clarity. Categorization is attempted but lacks clear reasoning and metrics.	No or inadequate application of k-fold cross-validation. Rudimentary model evaluation with significant issues. Rudimentary categorization with no clear reasoning or metrics.	understanding of these concepts.	
Comparative Performance Evaluation	**CLO2**	**4**	Exhibits exceptional mastery of integration and analysis.	Implements techniques to integrate and analyse all stages comparatively and effectively.	Achieves significant insights or illustrations wherever applicable.	Analysis provided is not insightful.	Analysis is performed but not in sequence.	No analysis attempted	**/20**
Hyper parameter tuning and Model interpretation	**CLO2**	**2**	Exceptional hyperparameter tuning with a clear documentation of tested hyperparameters, their rationale, and the optimization process. Profound model interpretation with detailed insights into influential features and precise predictions. Demonstrates a deep understanding of the model's behavior.	Proficient hyperparameter tuning with a clear rationale and effective optimization process. Proficient model interpretation with clear insights into influential features and predictions. Demonstrates a good understanding of the model's behavior..	Competent hyperparameter tuning with minor issues in documentation and rationale. Competent model interpretation with clear insights into influential features but with minor omissions. Demonstrates a basic understanding of the model's behavior.	Hyperparameter tuning is attempted but contains significant errors, lacks clarity in rationale, or has inadequate documentation. Model interpretation is rudimentary and lacks clear insights into influential features. Demonstrates significant	Rudimentary hyperparameter tuning with major issues in documentation, rationale, and optimization. Rudimentary model interpretation with no insights into influential features. Demonstrates a lack of understanding of the model's behavior.	No hyperparameter tuning or model interpretation attempted, or a complete lack of understanding of these concepts.	**/10**

						misunderstandings of the model's behavior.			
Documentation	**CLO2**	**4**	Provides a well-organized and clearly documented report with comprehensive explanations.	Presents a well-organized report with adequate documentation.	Presents a report with some organization but lacking in documentation.	Presents a disorganized report with minimal documentation.	Fails to present a coherent report or document the process.	No presentation or documentation.	**/20**
Total									**/100**

Any work submitted after the deadline's time will be subject to the standard University late penalties unless an extension has been granted, in writing by the Senior Tutor, in advance of the deadline. Details on the University's late penalties can be found here:

- https://www.southampton.ac.uk/~assets/doc/quality-handbook/Late%20Submission.pdf