School of London

COM7039M Machine Learning Assignment Brief

Contents

Module Details

Module code:	COM7039M	**Level of Study:**	7
Module Leader(s):	Dr. Rebecca Jeyavadhanam. B	**Credits:**	15
Assessment format:	Creative Artefact- A practical project to design and develop an ML prediction model with supporting documentation	**Method of submission:**	Turnitin within Moodle
Deadline or Assessment Period:	28th Jan 2025, 12Noon	**Feedback date and place:**	19th Feb 2025, Written feedback within Turnitin/Moodle
Assessment limits: length, load, word count, etc.	N/A	**Component number:**	1 of 1
Is this exempt from anonymous marking under the policy?	No	**Component weighting:**	100%

Assignment Description

This coursework aims to demonstrate the students' comprehensive understanding and knowledge of the Machine Learning module by evaluating their analytical abilities and strengths. It comprises two tasks designed to assess and challenge their analytical skills. These tasks have been carefully crafted to test the student's proficiency in applying the concepts learned throughout the module and to showcase their ability to tackle real-world scenarios and problems. Successful completion of these tasks will reflect the students' mastery of the subject matter and their capacity for critical and creative thinking.

The Assessment Consists of TWO Tasks:

In Task 1, you will be presented with a set of questions that require critical evaluation of your subject knowledge and understanding of the concepts related to the Machine Learning (ML) process and techniques.

In Task 2, you will be provided a programming exercise with a dataset to analyse using the Machine Learning approach. Your objective is to use suitable Machine Learning algorithms to predict and evaluate a model's accuracy. The dataset will contain various features and a target variable that you need to predict. You are expected to develop a model, followed by training, testing, and evaluating its performance. Your goal is to identify and highlight the best predictive model for classification. This includes comparing different models and selecting the one that offers the highest accuracy and best performance metrics for the classification task.

The content of this assignment must be supported by the inclusion of pertinent academic theories, concepts, models, and contemporary industrial insights. Provide a detailed and relevant description of your code. Ensure your work is accurately cited and referenced using the York St John Harvard Referencing Style.

Task 01: Theory Exercise – (20 Marks)

a. Discuss the K-means clustering algorithm in detail, including its working principles, advantages, disadvantages, and real-world applications. **(10 Marks)**

b. How would you evaluate the performance of the classification algorithm with appropriate metrics? **(10 Marks)**

Task 02: Programming Exercise -– (80 Marks)

a. Develop a classification model with Machine Learning techniques to detect hate speech from the Twitter dataset.

Dataset Description:

Assignment Description

The "Hate Speech and Offensive Language" dataset is collected from Twitter. It is primarily designed to support research and development in detecting and analyzing hate speech and offensive language on social media, distinguishing them from ordinary slang and neutral content.

Dataset Link:

https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset

Alternative Source:

The datasets are available to download from the Machine Learning (ML) module in the Moodle platform.

Guidelines to Prepare Your Assignment:

1. **Data Exploration and Pre-processing: (10 Marks)**

✓ Load and explore the dataset to gain insights into the data's characteristics.
✓ Handle missing values, if any, and perform data cleansing as required.
✓ Perform data visualization to understand variables' distribution and relationships with them.

2. **Feature Engineering Process: (15 marks)**

✓ Extract meaningful information from the dataset using Feature Extraction Technique
✓ Reduce the number of features using PCA as the Dimensionality Reduction Technique
✓ Select the most prominent features to capture the complex relationships
✓ Handle the outliers to prevent them from significantly affecting the model

3. **Model Selection and Training: (10 Marks)**

✓ Split the dataset into training and testing sets.
✓ Select appropriate machine learning algorithms (e.g., logistic regression, decision trees, random forests, support vector machines, etc.) for the predictive modeling task.
✓ Train the selected models on the training data and evaluate their performance on the testing data.

4. **Hyperparameter Tuning: (10 Marks)**

✓ Fine-tune the hyperparameters of the chosen algorithm to optimize the model's performance and avoid model overfitting.

5. **Model Evaluation: (20 Marks)**

✓ Compare the performance of different models using appropriate evaluation metrics such as accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (AUC-ROC), Confusion matrix, and Logarithmic loss (Log Loss).
✓ Comparative analysis to identify the best-performing model for classification.

6. **Model Deployment: (10 Marks)**

Assignment Description

✓ Deploy the model to generate predictions for new and previously unseen data.

7. Conclusion and Recommendations: (5 marks)

✓ Summarize the key findings of your analysis, highlighting the model's performance and any insights gained.

Propose potential improvements or additional steps that could be taken to enhance the system.

Learning Outcomes

PLOs 7.1-7.7

7.1Evaluate computer science concepts and principles and their application to the effective

design, implementation, and usability of computer-based systems.

7.2 Apply the findings of advanced scholarship and/or contemporary research and practice to

the solution to computer science problems

7.3 Critically evaluate computer science problems, including those at the forefront of the field.

7.4 Demonstrate operation within applicable professional, legal, social, and ethical frameworks.

7.5 Demonstrate originality and creativity in the solution of computer science problems.

7.6 Recommend, with detailed justification, the appropriate computer science principles and

practices to apply to significant domain-specific activity.

7.7 Apply standards, quality processes, and engineering principles to the solution of computer science problems.

Advice and Guidance

Submission Guidelines:

Advice and Guidance

- Prepare a comprehensive report documenting your approach, methodologies, results, and insights gained from the project.

- Include code snippets, visualizations, and explanations to support your findings. Your report should be clear, concise, and well-organized.

- **Submit the complete report, including both Task 1 and Task 2, as a single PDF file. The code must be submitted as a supporting document in the .ipynb file format.**

Additional Guidelines for Students:

Students must submit their own work. They must acknowledge the sources used in this assignment, failure to acknowledge would be plagiarism which is an academic offence and a penalty can be imposed. Students need to write by reading other papers on their own with citations and leave references at the end of the assignment.

Students work would be submitted to the national plagiarism facility. This identifies the sources from the internet and other extensive databases. Once the student's work has been submitted to detection services, work is stored in databases electronically and compared their work from other sources. It is necessary to keep a backup of their work. Students' materials would be stored in the database electronically for indefinite periods.

It is essential that you acknowledge the source of any research, information, ideas, opinions, theories, or other material which is not your own. Effective referencing, quoting, paraphrasing, and summarising show evidence of the reading you have done and ensure that you avoid accusations of plagiarism.

The University's fundamental stance on the use of Turnitin is geared toward supporting students' academic development. You can use this link to check your work for areas where you might be at risk of plagiarising.

Please submit your assignment on time. All assignments may be electronically submitted using Turnitin (via Moodle) by midnight on the due date. Please do not submit your assignment last minute. Please also allow time for any problems or issues with systems.

The work you present should be your own work, and not just copied from others. You can quote from others, but you must say who the author is and use quotation marks or paraphrase. If you do not do so, we will investigate your work for academic misconduct. This is particularly likely if your Turnitin similarity score is above 25% and/or individual matches are above 6%.

Advice and Guidance

If you require support with your study skills, please visit https://www.yorksj.ac.uk/students/study-skills/

Please refer to the York St John University Code of Practice for Assessment and Academic Related Matters 2024-25.

We ask that you pay particular attention to the academic misconduct policy. Penalties will be applied where a student is found guilty of academic and/or ethical misconduct, including termination of programme (**Policy Link**).

You are required to keep to the word limit set for an assessment and to note that you may be subject to penalty if you exceed that limit. You are required to provide an accurate word count on the cover sheet for each piece of work you submit (**Policy Link**).

For late or non-submission of work by the published deadline or an approved extended deadline, a mark of 0NS will be recorded. Where a re-assessment opportunity exists, a student will normally be permitted only one attempt to be re-assessed for a capped mark (**Policy Link**).

An extension to the published deadline may be granted to an individual student if they meet the eligibility criteria of the (**Policy Link**).

How is this assessment marked?

Your work will be marked according to the assessment instructions provided within this document and the selected Learning Outcomes' (LOs) (see above).

Furthermore, this assessment is marked using the assessment marking criteria or a similar rubric that aligns with the University's Generic Assessment Descriptors (see below).[1] This is to ensure all assessment decisions are comparable regardless of the discipline or mode of assessment.

Please note that you ***must*** *meet the required baseline standards (50 – 59%) which will include the LOs and minimum expectations of the assessment. Further still, you must ensure you meet the requirements of each grade boundary to progress to the next, i.e., you should demonstrate your learning through the standards of the Pass, Merit and Distinction to reach a Distinction (70 – 84%). These standards are designed to scaffold and build your learning to achieve your fullest potential in each criterion being assessed.*

Deliverables for Task 1 and Task 2

		Deliverables	Marks
Task 01- 20 Marks	a	An extraordinary conceptual understanding of K-means clustering algorithm, advantages, and disadvantages with real-world applications. If any examples are provided.	**10 Marks**
	b	An appropriate description of all the metrics like accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (AUC-ROC), Confusion matrix, and Logarithmic loss (Log Loss.)	**10 Marks**
Task 02- **80 Marks**	**Data Exploration and Pre-processing**	Correct handling of missing values, outliers, and data normalization. Effective exploratory data analysis (EDA) to gain insights into the dataset. Provide a well-documented Jupyter Notebook or script containing the code for data preprocessing steps. Ensure that each step is properly commented on to explain its purpose and functionality.	**10 Marks**
	Feature Engineering	Present code segments that generate new features based on domain knowledge or creative insights. Discuss the encoding methods chosen and their suitability for the problem. Applying mathematical transformations to numerical features. Scaling numerical features to ensure they are on similar scales. Discuss the feature selection and its importance.	**15 Marks**
	Model Selection and Training	Correct implementation of selected machine learning algorithms should be presented. Splitting the dataset into training and testing (validation) sets using techniques like the train-test split or k-fold cross-validation. Adequate use of libraries and tools to streamline the implementation process.	**10 Marks**

	Hyperparameter Tuning	Explanation of the method you used for hyperparameter tuning reasons for selecting this method and how it suits your specific problem. Data Splitting Strategy: How you divided your data into training, validation, and test sets. Model Training and Evaluation Protocol: Description of how you trained and evaluated models for different hyperparameter configurations. Explanation of the performance metric(s) you used to assess model performance.	**10 Marks**
	Model Evaluation	Accurate evaluation of model performance using relevant metrics (e.g., accuracy, precision, recall, F1-score). Comprehensive comparison of multiple models and their strengths/limitations. Insightful interpretation of results and trends observed.	**20 Marks**
	Model Deployment	Testing the developed model using real-world data or unseen data that it performs as expected and provides accurate predictions and validates the model performance.	**10 Marks**
	Conclusion and Recommendations	The clear and organized structure of the report with proper sections (Introduction, Methodology, Results, Discussion, Conclusion). Coherent explanations of the implemented algorithms and techniques. Effective visualization of results through graphs, charts, and tables. Cohesive and well-written analysis of findings and conclusions.	**5 Marks**
		Total Marks	**100 Marks**

Marking Criteria

Pass Grade Bands (100 – 50) (Learning Outcomes must be met)
Fail Grade Bands (49 – 0) (Learning Outcomes are not met)

Assessment Criteria		Pass (50 – 59)	Merit (60 – 69)	Distinction (70 – 84)	Distinction (85 – 100)	Borderline Fail (45 - 49) **(Credits may be compensated)**	Fail (30 - 44) **(Credits may not be compensated)**	Fail (0 - 29) **(Credits may not be compensated)**
Task 1a-10%	**An extraordinary conceptual understanding of the Naïve Bayes algorithm, including its advantages, disadvantages, and real-world applications.**	Demonstrates a deep and insightful understanding of the NB algorithm with detailed examples and critical analysis.	Shows a strong understanding with relevant examples and solid analysis.	Provides a good understanding with some examples and analysis.	Demonstrates adequate understanding with basic examples and some discussion.	Shows limited understanding with insufficient examples or analysis.	Demonstrates a poor understanding with little to no relevant examples.	Fails to demonstrate understanding of algorithm.
Task 1b-10%	**An innovative approach to problem-solving with creative insights and solutions.**	Exhibits exceptional creativity and originality in problem-solving.	Shows strong creativity with effective problem-solving approaches.	Demonstrates good creativity with some innovative solutions.	Provides adequate creativity with basic problem-solving approaches.	Limited creativity with few innovative solutions.	Shows minimal creativity with ineffective problem-solving.	Fails to demonstrate creativity or effective problem-solving.

Data Exploration and Pre-processing-10%	Excellent handling of missing values, outliers, and data normalization. Effective and insightful EDA with a well-documented Jupyter Notebook or script.	Strong handling of data issues and effective EDA. Well-documented with minor gaps in explanation.	Good handling of data issues and EDA. Adequate documentation and explanation.	Adequate handling of data issues with basic EDA. Documentation and explanations are present but may lack depth.	Limited handling of data issues or EDA. Incomplete documentation or explanation.	Poor handling of data issues with insufficient EDA. Inadequate documentation and explanations.	Fails to handle data issues effectively. Lacks proper EDA and documentation
Feature Engineering Process-15%	Innovative and effective feature engineering with detailed explanation of encoding methods, mathematical transformations, scaling, and feature selection.	Strong feature engineering with good explanation of methods and transformations.	Good feature engineering with some explanation of methods and transformations.	Adequate feature engineering with basic explanation of methods and transformations.	Limited feature engineering with minimal explanation.	Poor feature engineering with insufficient explanation.	Fails to demonstrate effective feature engineering.
Model Selection and Training-10%	Excellent implementation of algorithms with thoughtful data splitting and optimal use of tools and libraries.	Strong implementation with appropriate data splitting and effective use of tools.	Good implementation with correct data splitting and adequate tool usage.	Adequate implementation with basic data splitting and tool usage.	Limited implementation with inappropriate data splitting or tool usage.	Poor implementation with ineffective data splitting or minimal tool usage.	Fails to implement models correctly or use tools effectively.
Hyperparameter Tuning-10%	Comprehensive evaluation using relevant metrics with insightful comparison and interpretation of results.	Accurate evaluation with good comparison and interpretation of multiple models.	Good evaluation with appropriate metrics and some comparison of models.	Adequate evaluation with basic metrics and limited comparison of models.	Limited evaluation with minimal use of metrics and comparison.	Poor evaluation with inadequate metrics and no comparison of models.	Fails to evaluate models effectively or provide meaningful insights.
Model Evaluation-20%	Comprehensive evaluation using relevant metrics with insightful comparison and interpretation of results.	Accurate evaluation with good comparison and interpretation of multiple models.	Good evaluation with appropriate metrics and some comparison of models.	Adequate evaluation with basic metrics and limited comparison of models.	Limited evaluation with minimal use of metrics and comparison.	Poor evaluation with inadequate metrics and no comparison of models.	Fails to evaluate models effectively or provide meaningful insights.
Model Deployment-10%	Thorough testing of the model with real-world or unseen data, demonstrating accurate predictions and validation of performance.	Effective testing with real-world or unseen data and validation of performance.	Good testing with some validation of performance using unseen data.	Adequate testing with basic validation of performance.	Limited testing with insufficient validation of performance.	Poor testing with minimal or ineffective validation of performance.	Fails to test or validate model performance effectively.
Report Structure and Clarity-5% [Communication]	Clear, organized report with detailed sections, insightful explanations, and effective visualizations. Cohesive analysis and conclusions.	Well-structured report with good explanations and visualizations.	Clear report with adequate structure, explanations, and some visualizations.	Adequate report with basic structure and analysis, though visualizations may be lacking.	Limited report with unclear structure and insufficient analysis or visualizations.	Poor report with minimal organization, analysis, and visualizations.	Fails to provide a coherent report or meaningful analysis and

							visualizations .