

TEMASEK POLYTECHNIC
SCHOOL OF INFORMATICS & IT**Diploma in Information Technology**
Machine Learning for Developers (CAI2C08)
AY2025/2026 October Semester
Project Specifications**Background**

In today's data-driven world, technical skills alone are not enough. Machine learning (ML) developers must do more than build models — they need to effectively communicate solutions with stakeholders across business units, justify decisions with evidence, and adapt to changing business needs.

This project simulates a real-world scenario to prepare you for your Major Project, Student Internship Programme in Year 3, and future industry roles. You will take on the role of a founder in a start-up company, where you will **develop** and convincingly **present** your machine learning solution (web app).

Objectives

By completing this project, you will:

- identify problems suitable for machine learning
- perform exploratory data analysis to data using Python libraries
- apply machine learning concepts and methods to write solutions using Python
- deploy solutions to an online application platform using Python and Streamlit
- communicate your results clearly and respond to business challenges.

Project Components & Deadlines

- This project contributes **70% to your overall subject grade** for the module, with a total mark of 70 allocated to this component.
- This is an **INDIVIDUAL** project and it consists of two main components:

Deliverables	Percentage	Submission	Due Date
(1) Program Code	50%	TP-LMS	11 Feb (Wed) 9.00am
(2) Project Demo and Presentation	20%	TP-LMS In classroom (Business attire)	Slides: 25 Feb (Wed) 9.00am Demo and Presentation (Pitching): 26 Feb (Thu) OR 27 Feb (Fri)

Project Task and Scope

Imagine yourself as the founder of a start-up company.

Your mission is to use machine learning to solve a real-world problem and convince potential investor (your tutor) of your solution.

The following tasks are required:

1. **Identify a practical problem** that can be addressed using **supervised machine learning**.
 - Choose a topic that interests you and it is a practical real-world problem in domains such as finance, healthcare, marketing, logistics, services, manufacturing, sustainability, and cybersecurity.
2. Source and **analyze data**:
 - Find a public dataset with **at least 1,000 samples** (rows) from one of these sources, but not limited to:
 - UCI Machine Learning Repository <https://archive.ics.uci.edu/ml/index.php>
 - Kaggle Datasets <https://www.kaggle.com>
 - HuggingFace Datasets <https://huggingface.co>
 - Singapore Open Data Portal <https://data.gov.sg>
 - Google Dataset Search <https://datasetsearch.research.google.com>
 - Your dataset should contain **multiple columns** for feature engineering or selection.
 - Identify the column(s) you are predicting and correctly determine whether the problem is a **regression** or **classification** task.
 - **Do NOT** use any dataset that contains unstructured data (e.g., paragraphs of text) or images or solely time series data (e.g., cryptocurrency or stock prices)
 - Perform exploratory data analysis (EDA) to understand the data and its relevance to your chosen problem.
 - Ensure your dataset is **not synthetically generated**.
 - **Seek your tutor's approval** (during Project Checkpoint 1) before developing your ML solution as some datasets may not be ideal for machine learning projects.
 - Ensure that the scope of work required by the choice of your dataset can allow you to complete your project by the specified deadline.
3. **Develop your machine learning solution** using Python 3:
 - Implement models using the **scikit-learn library ONLY**. **Do NOT** use any other libraries for model implementation.
 - Document your code, including rationale for
 - data preprocessing and/or feature engineering
 - and algorithm/model selection.
 - Justify the choice of your evaluation metric(s).
 - **Experiment with different models** (including feature engineering and/or feature selection), and then tune the parameters and hyperparameters for the **best model** (use only RandomizedSearchCV, with no more than 3 values for each hyperparameter), and justify your choice of model based on evidence and business needs. State any assumption (if applicable) that you have made.
 - **You need to show all the steps used to train and evaluate your models, along with outputs** (e.g., graphs, logs).

4. Deploy your solution:

- Build a user-friendly web application (using Streamlit) that demonstrates your model's capabilities. The web application should be aesthetically appealing, user-friendly, and appropriate for the target audience.
- Prepare a Word document with the web app link and three screenshots showing its functionality.

5. Pitch your solution to investor:

- Create a concise and compelling pitch slide deck summarizing your web app solution: problem, key approach, and results / business impact.
- Present and demonstrate to a potential investor (lecturer) your web app solution as a compelling product that addresses the identified business problem.
- Respond to questions and a case challenge to demonstrate your ability in handling concerns, overcoming constraints, and assessing alternative solutions.

(1) Program Codes (50%)

- You need to develop your codes and documentation as outlined in Tasks 3 and 4 in *Project Task and Scope* (previous section).
- Please refer to the Marking Rubrics for more details.

Submission 1: Program Codes

Submit your project as a zipped file to POLITE Mall (TP-LMS) under the Assessment folder.

Your zipped file should include the following:

1. Jupyter notebook file (*.ipynb)
 - Code
 - Documentation (output, graphs, training logs, comments etc)
 - Analysis and Justification
2. Streamlit web application code file (*.py)
3. Word document (*.docx) – Use template given
 - Declaration of Originality
 - Source of dataset
 - GitHub link
 - Streamlit link
 - 3 screenshots
 - (1) Before any option is selected
 - (2) After an option is selected (should show a prediction)
 - (3) After another option is selected (should show a change in value for the prediction)
 - Screenshots of version control or development logs (if applicable)

Note:

- Your zipped folder should be named as *<TP-LMS Name>_<Student ID>.zip* (e.g., Aiman Tan_2401234A.zip). Do NOT compress in RAR format.
- Only the latest submission will be graded.
- If you are resubmitting, always compress ALL components again into a single *.zip file and upload it. Where possible, remove any previously uploaded files to avoid confusion.
- Upload your assignment early; do NOT wait till the last minute.
- When submitting assignments, each student is responsible for ensuring that they do NOT submit an empty zip file or a corrupted file.
- Students must also ensure they are using a suitable device when working on assignments / projects, maintain proper backups of their work and do not cite 'laptop issues' as a reason for assignment extensions as such reasons will NOT be considered.

(2) Demo and Presentation (20%)

- You will have up to **3 minutes** to deliver your pitch.
- Afterwards, you will respond to questions and a case challenge to demonstrate your ability in handling concerns, overcoming constraints, and assessing alternative solutions.
- **Please refer to the Marking Rubrics for more details.**
- To help you better prepare for your presentation, you may refer to MAPS framework here:
 - <https://bit.ly/MAPS-Framework-pdf>
 - <https://bit.ly/MAPS-Framework-vid>
- You can also use Microsoft Presenter Coach to practise and improve your pitch delivery. This tool provides feedback to improve your pace, intonation, and body language through real-time feedback, thus helping you prepare for a confident and effective presentation. More information can be found in TP-LMS.

Submission 2: Demo and Presentation Slides

Submit your slides (*.pptx) to POLITE Mall (TP-LMS) under the Assessment folder.

1. Slides

- You have a **maximum of 10 slides**, excluding cover page and references.
 - Be concise and only include the most important information.
- Suggested slide content:
 - (1) **Problem & Dataset**
 - What is the problem that you are resolving? Why does it matter?
 - What is your dataset? (source, relevance, size)
 - (2) **Web App Demonstration**
 - How does your deployed web application look like? (video or live demo)
 - What are the main features? E.g., how users interact with app, how the app presents the output.
 - (3) **Key Approach, Methods, Results & Impact**
 - What are the key results (e.g., accuracy)? How do these results inform or impact the business problem?
 - What machine learning method have you used?
 - What are the key data processing and/or feature engineering steps etc.?
 - Why have you chosen this approach?
 - (4) **Conclusion & Next Steps**
 - What is the value of your solution?
 - Why should your solution be considered?

(5) References

- List key sources and datasets.
- List of research materials and external sources you used for the project.
(Website links/ APA format for research papers).

Guidelines on using Gen AI for your project

- You may use Gen AI tools to generate ideas and prepare your pitch.
- Do not copy and paste exact results, as AI can be unpredictable, and they may not always be correct.
- Verify the results (e.g., codes, suggestions, and explanation).
- You are responsible for your own work.
- Cite where you have used Gen AI tools in the Declaration of Originality.
- Refer to the ChatGPT Playbook Student Edition in POLITE Mall (TP-LMS) for guidelines on what you can and CANNOT do.

Policies & Penalties for Late Submission

late and < 1 day	10% deduction from absolute mark given for the assignment (e.g., 75 marks (100 marks max) becomes 65 marks (deduct 10% of 100 marks))
late >= 1 and < 2 days	20% deduction from absolute mark
late >= 2 days	No marks awarded

- Note that "day" includes **non-working days** (Saturday, Sunday, and public holidays).
- General MC/LOA is **NOT** considered as valid reason for extension of assignment submission.

Penalty for Plagiarism

Academic-related Offences	Outcome
<p>Type of offence</p> <ul style="list-style-type: none"> • Cheating on Semestral Examination (including possession of unauthorised materials as verified by the Examiner as relevant to the examination paper) • Cheating on Coursework such as projects, term tests and assessments • Plagiarism Please refer to the Plagiarism Policy <p>(For plagiarism, if two or more students are involved in plagiarism, the same penalty shall apply to all students, including the student who knowingly allowed his or her work to be copied and group members, if applicable.)</p>	<p>Punitive actions could include the following:</p> <ul style="list-style-type: none"> • Warning letter • Failing the assessment item / component (i.e. '0' mark, 'DC' grade) - if assessment item is less than 30% of subject assessment • Fail subject (i.e. '0' mark, 'DC' grade) - if assessment item is more than or equal to 30% of subject assessment • Fail all subjects • Suspension from course of study for a minimum period of one semester • Dismissal from the Polytechnic <p>Plagiarized work will be reported and awarded zero marks.</p>

Marking Rubrics
(1) Program Codes (50%)

Component	Evaluation Criteria	A (≥ 80%)	B (70 to <80%)	C (60 to <70%)	D (50 to <60%)	F (<50%)
Code Explanation & Documentation (8%)	Code Clarity & Documentation (4%)	<ul style="list-style-type: none"> Code is organized and readable Clearly document all key steps and why each step is needed with meaningful comments 	<ul style="list-style-type: none"> Code is mostly organized Describe most key steps with brief rationale 	<ul style="list-style-type: none"> Code is mostly organized List steps but rationale is minimal or missing for more than 1 major step 	<ul style="list-style-type: none"> Code has some readability issues (e.g., inconsistent naming) Justification or comments largely absent or unhelpful 	<ul style="list-style-type: none"> Code is unreadable Lack comments
	Version Control / Development Logs (4%)	<ul style="list-style-type: none"> Clear, observable iterative development (e.g., logs / commit history) 	<ul style="list-style-type: none"> Observable evidence of iterative development 	<ul style="list-style-type: none"> Minimal evidence of iterative development, such as a single mention or figure indicating change, but lacks comparison or detail. 	<ul style="list-style-type: none"> Evidence of iterative development is ambiguous or unclear; may be implied but not explicitly shown 	<ul style="list-style-type: none"> Present only final result, with no explicit evidence of iterative development
Iterative Model Development: Creating and Evaluating the Model (32%)	Exploratory Data Analysis (EDA) & Interpretation (8%)	<ul style="list-style-type: none"> Present insightful EDA with relevant use of visualisations (e.g., distribution, correlation/feature importance, target vs key features) that are clearly labelled Interpret trends and/or outliers and state implications for modelling 	<ul style="list-style-type: none"> Present a few relevant visualisations Interpret trends and/or outliers 	<ul style="list-style-type: none"> Present one visualisation or only summary statistics Limited interpretation; implications for modelling not stated 	<ul style="list-style-type: none"> Present one visualisation or only summary statistics Unlabelled / irrelevant or no interpretation 	<ul style="list-style-type: none"> No EDA
	Model Diversity, Comparison & Selection Rationale (8%)	<ul style="list-style-type: none"> Train two or more distinct algorithms with clear output logs shown Report test set results with baseline comparison Present quantitative and qualitative comparison of model performance Provide a clear rationale for model selection based on performance, interpretability, or deployment needs 	<ul style="list-style-type: none"> Train two or more models Report test set results with baseline comparison Present basic comparison Provide a brief rationale for selection, possibly lacking in depth or clarity 	<ul style="list-style-type: none"> Train two or more models Report test set results with no baseline comparison OR Present some comparison, but no rationale for selection 	<ul style="list-style-type: none"> Train only one model Report test set results No rationale for selection 	<ul style="list-style-type: none"> No working model

Component	Evaluation Criteria	A (≥ 80%)	B (70 to <80%)	C (60 to <70%)	D (50 to <60%)	F (<50%)
(Continued) Iterative Model Development: Creating and Evaluating the Model (32%)	Feature Engineering / Feature Selection with Rationale (8%)	<ul style="list-style-type: none"> • Apply meaningful feature engineering and/or selection • Show measurable improvement • Explain changes 	<ul style="list-style-type: none"> • Apply meaningful feature engineering and/or selection • Show minor measurable improvement • Explanation is brief 	<ul style="list-style-type: none"> • Apply meaningful feature engineering and/or selection • No measurable improvement • Generic explanation that does not tie to data/metrics 	<ul style="list-style-type: none"> • Present irrelevant or unjustified changes • No measurable improvement 	<ul style="list-style-type: none"> • No feature engineering or selection <p>OR</p> <ul style="list-style-type: none"> • No attempt at model refinement or changes made are irrelevant or unjustified
	Hyperparameter Tuning (4%)	<ul style="list-style-type: none"> • Apply tuning (vary two hyperparameters) • Present detailed comparison logs or tables • Explain impact and select final settings based on documented improvement VS baseline • Use reproducible setup 	<ul style="list-style-type: none"> • Vary two hyperparameters • Present comparison • Provide brief explanation of impact for the final setting 	<ul style="list-style-type: none"> • Change only one hyperparameter • Present partial comparison (e.g., missing test score) • No clear explanation of impact 	<ul style="list-style-type: none"> • Change only one hyperparameter • No comparison • No clear explanation of impact 	<ul style="list-style-type: none"> • No tuning attempted
	Evaluation Metric Selection, Interpretation & Justification (4%)	<ul style="list-style-type: none"> • Use task-appropriate metric • Present clear link to the business outcome (e.g., a few sentences connecting metric to impact). 	<ul style="list-style-type: none"> • Use task-appropriate metric • Present brief rationale with limited link to business/context 	<ul style="list-style-type: none"> • Use task-appropriate metric • Present generic rationale and not tied to the business/context 	<ul style="list-style-type: none"> • Use inappropriate metric • Rationale is weak or contradictory 	<ul style="list-style-type: none"> • Use incorrect or inconsistent metrics • Rationale not explained

Component	Evaluation Criteria	A (≥ 80%)	B (70 to <80%)	C (60 to <70%)	D (50 to <60%)	F (<50%)
Deployment of Model (10%)	Correct prediction of output without error (4%)	<ul style="list-style-type: none"> Web app produces correct outputs for all tested cases No errors or crashes Include input validation and user-facing error message(s) 	<ul style="list-style-type: none"> Web app produces correct outputs in all tested cases Minor warnings that do not affect outputs 	<ul style="list-style-type: none"> Web app produces correct outputs for most tested cases Occasional errors or crashes 	<ul style="list-style-type: none"> Web app produces correct outputs for some tested cases Frequent errors or crashes 	<ul style="list-style-type: none"> Web app fails to produce correct outputs Persistent errors or crashes Unusable for intended purpose
	Interactive application (6%)	<ul style="list-style-type: none"> Highly interactive, responsive, and user-friendly app All interactive elements (buttons, forms, feedback) work smoothly and enhance user experience Design and visuals are appealing, and labels are suitable for target audience 	<ul style="list-style-type: none"> Interactive, responsive, and user-friendly app Design and visuals are consistent (e.g., at most 2 minor issues), and labels are mostly clear (e.g., at most 2 vague terms) for target audience 	<ul style="list-style-type: none"> Interactive, responsive, and user-friendly app Some elements may not respond as expected; user experience is acceptable but not engaging 	<ul style="list-style-type: none"> Largely static interface 	<ul style="list-style-type: none"> No interactivity Present many broken elements that do not work or respond poorly; user experience is frustrating

(2) Demo and Presentation (20%)

Component	Evaluation Criteria	A (≥ 80%)	B (70 to <80%)	C (60 to <70%)	D (50 to <60%)	F (<50%)
Pitching Problem, Solution & Results (12%)	Slide Organization & Visuals (4%)	<ul style="list-style-type: none"> Kept to slides limit Clear, legible, well-organized slides Informative visuals (charts, tables, diagrams) with supporting explanation 	<ul style="list-style-type: none"> Kept to slides limit Mostly clear and legible slides Mostly clear and informative visuals 	<ul style="list-style-type: none"> Adequate slides Visuals lack labels or insufficiently explained 	<ul style="list-style-type: none"> Cluttered or hard-to-follow slides Visuals lack labels or insufficiently explained 	<ul style="list-style-type: none"> No slides or present unreadable content
	Presentation Delivery & Preparedness (4%)	<ul style="list-style-type: none"> Kept strictly to time Delivery is clear, confident, and engaging Presenter maintains good eye contact, and uses appropriate body language for a professional presentation. Presenter is well-prepared, speaks audibly, and dressed in appropriate attire 	<ul style="list-style-type: none"> Kept strictly to time Delivery is mostly understandable and clear Presenter shows generally professional eye contact/body language. Presenter is prepared, speaks audibly, and dressed in appropriate attire 	<ul style="list-style-type: none"> Delivery is mostly understandable and clear Presenter reads off slides frequently, speaks unclearly or too softly 	<ul style="list-style-type: none"> Delivery is poor or presenter reads off slides all the time Presenter appears unprepared, speaks unclearly or too softly 	<ul style="list-style-type: none"> Delivery is unclear Presenter is unprepared
	Solution Alignment & Convincing Justification (4%)	<ul style="list-style-type: none"> Address stated problem and link outcomes to business value Justification is convincing, relevant, and supported by evidence (metrics/tables) 	<ul style="list-style-type: none"> Address stated problem and link outcomes to business value Justification is largely convincing, relevant, and includes some evidence 	<ul style="list-style-type: none"> Addresses the stated problem, but justification is limited to basic description or restatement of results Justification includes minimal evidence without context or comparison (e.g., no baseline) 	<ul style="list-style-type: none"> Partially address the problem (e.g., evaluate a different outcome than stated) Justification is weak or unsupported (claims made without relevant evidence, or evidence does not relate to the problem) 	<ul style="list-style-type: none"> Solution does not address the problem Justification is missing or irrelevant or entirely unsupported No evidence provided, or evidence is unrelated to the solution
Response to Case and Questions (8%)	Solution Comparison & Technical Justification (8%)	<ul style="list-style-type: none"> Provide insightful, critical comparison of solution and alternatives Provide clear and technically accurate justification 	<ul style="list-style-type: none"> Provide clear comparison Provide generally accurate justification 	<ul style="list-style-type: none"> Provide minimal comparison Provide generally accurate justification Handle questions correctly, but with follow-up prompts 	<ul style="list-style-type: none"> Provide unclear comparison Provide unclear or generic justification Struggle with deeper questions or prompts 	<ul style="list-style-type: none"> Unable to justify or present irrelevant discussion Did not answer, avoid or deflect questions

(: End of Project Specifications :)