Xi'an Jiaotong-Liverpool University
西交利物浦大學

XJTLU Entrepreneur College (Taicang) Cover Sheet

Module code	DTS304TC: Machine Learning	**School title**	School of AI and Advanced Computing
Assessment title	Coursework Task 1	**Assessment type**	Coursework
Submission deadline	01/May/2026 23:59		

I certify that I have read and understood the University's Policy for dealing with Plagiarism, Collusion and the Fabrication of Data (available on Learning Mall Online).

My work does not contain any instances of plagiarism and/or collusion.

My work does not contain any fabricated data.

By uploading my assignment onto Learning Mall Online, I formally declare that all of the above information is true to the best of my knowledge and belief.

Scoring – For Tutor Use	
Student ID	
Theory and Reflection PDF Word Count (Filled by Students)	

Stage of Marking	**Marker Code**	**Learning Outcomes Achieved (F/P/M/D) (please modify as appropriate)**			**Final Score**
		A	**B**	**C**	
1st Marker – red pen					
Moderation – green pen	**IM Initials**	The original mark has been accepted by the moderator (please circle as appropriate):			Y / N
		Data entry and score calculation have been checked by another tutor (please circle):			Y
2nd Marker if needed – green pen					

For Academic Office Use			**Possible Academic Infringement (please tick as appropriate)**	
Date Received	**Days late**	**Late Penalty**	☐ **Category A**	Total Academic Infringement Penalty (A,B, C, D, E, Please modify where necessary) _____
			☐ **Category B**	
			☐ **Category C**	
			☐ **Category D**	
			☐ **Category E**	

DTS304TC Machine Learning
Coursework - Assessment Task 1

- Percentage in final mark: 50%
- Assessment type: individual coursework
- Submission files: one Jupyter notebook (.ipynb), one Coursework Answer Sheet / Theory and Reflection PDF, and one hidden-test CSV

Learning outcomes assessed

- A. Demonstrate a solid understanding of the theoretical issues related to problems that machine-learning methods try to address.
- B. Demonstrate understanding of the properties of existing machine-learning algorithms and how they behave on practical data.

Notes

- Please read the coursework instructions and requirements carefully. Not following these instructions and requirements may result in a loss of marks.
- The formal procedure for submitting coursework at XJTLU is strictly followed. Submission link on Learning Mall will be provided in due course. The submission timestamp on Learning Mall will be used to check late submission.
- 5% of the total marks available for the assessment shall be deducted from the assessment mark for each working day after the submission date, up to a maximum of five working days.
- All modelling work must be completed individually. Discussion of general ideas is allowed, but code, experiments, and notebooks must be independently developed.
- You may not use ChatGPT to directly generate answers for the coursework. High-scoring work must demonstrate your own experimental design, controlled comparisons, failure analysis, and image-level interpretation. ChatGPT or similar tools may be used only in a limited support role such as code understanding, debugging, or grammar support. They must not replace your method design, ablation logic, qualitative analysis, or reflection. Generic AI-produced descriptions without matching evidence in code, tables, figures, and discussion will not receive high marks.
- If you use AI tools or outside code in any meaningful way, you must fully understand, verify, and take ownership of every method, number, figure, and written claim that appears in your submission.

Question 1: Notebook-Based Coding Exercise - Insurance Premium-Risk Classification (60 Marks)

In this coursework you will build and improve a multiclass classifier for a fictionalised health-insurance dataset. The task is to predict whether each applicant belongs to a Low, Standard, or High premium-risk group before pricing a policy. The dataset is intentionally realistic: it mixes numerical and categorical variables, contains missing values and dirty entries, and includes some fields that require careful handling to avoid weak modelling practice or label leakage.

Your work should show a clear machine-learning workflow: build a sensible first pipeline, compare model families, apply stronger hyperparameter optimisation, complete one compulsory improvement category plus at least one optional category, carry out a compact K-Means/Gaussian Mixture Model (GMM) exploration, and then produce a hidden-test CSV using validation evidence only.

The prediction target variable is ‘premium_risk’, and it has 3 imbalanced classes: Standard, High, Low. The dataset contains 33 raw columns: admin/PII columns, synthetic noise features, 1 leakage feature, and genuine predictors.

Unless otherwise stated, macro-F1 is the primary validation metric because the dataset is imbalanced; accuracy is reported as a secondary metric.

(A) Clean First Pipeline and Baseline Modelling (8 marks)

- Load the provided training and validation files and define a consistent target / feature setup.
- Handle leakage features, dirty values, missing values, and categorical variables sensibly. A compact sanity check is enough; a long data-audit section is not required.
  Important: The dataset contains a leakage feature. You must identify and remove it before proceeding to the next stage of analysis; otherwise, the classification results will be severely biased by this leakage and will not be meaningful. If this occurs, multiple parts of your Coursework 1 may be affected, which could significantly impact your marks.
- Build one baseline modelling pipeline.
- Report at least one validation result using accuracy and macro-F1 score and include a confusion matrix for the baseline model.
- Keep preprocessing consistent across train, validation, and hidden-test files.

(B) Controlled Comparison: Random Forest and One Boosting Model (8 marks)

- Using the same preprocessing pipeline, validation split, and evaluation metric (primary metric is macro-F1 also report accuracy), carry out an initial controlled comparison between one Random Forest model and one boosting model.
- Default XGBoost is recommended because it provides a richer tuning space later, but others may also be used. Default settings or only light sensible adjustments are acceptable in this section.
- In the notebook, report the validation result of each model and support the comparison with one or two additional analyses, such as class-wise metrics, a confusion matrix, train-versus-validation behaviour, or stability / sensitivity after tuning.
- Your goal is not to prove that one model type always wins. Your goal is to compare the two models fairly, explain the high-level learning difference between bagging and boosting, and use your own notebook evidence to give a careful, dataset-specific interpretation. A generic textbook answer without reference to your own results will receive limited credit.

(C) Advanced Hyperparameter Optimisation (12 marks)

- At least one main model should be tuned with a genuinely advanced strategy such as Optuna/TPE, Bayesian optimisation, Hyperopt, Ray Tune, or another comparably strong approach.
- Hyperparameter tuning should optimise macro-F1 score on the validation set, and the final tuned result should be reported using both accuracy and macro-F1.
- RandomizedSearchCV alone is normally not enough for the top band.
- Explain briefly why your search space and optimiser are reasonable for the chosen model.

(D) Personalised Improvement Work (18 marks)

You must complete one compulsory category based on the last digit of your XJTLU student ID, plus at least one additional optional category of your choice. A second optional category is recommended for stronger differentiation but is not compulsory. You should report accuracy and macro-F1 for improved models and include class-wise metrics where helpful. A compact ablation table should normally be included in the notebook for the personalized improvement work

Last digit	Compulsory category
0-1	Category A - Data quality and missingness
2-3	Category B - Feature representation and engineering
4-5	Category C - Imbalance and objective design
6-7	Category D - Model robustness, calibration, or ensembling
8-9	Category E - Fairness, diagnostics, or interpretability

Category	Examples of what may be done	What good evidence looks like
A	better missing-value strategy; MissForest or iterative imputation; sensible outlier handling; value cleaning	A concise before/after comparison with a short explanation of why the data handling changed the result
B	feature crosses; grouped categories; alternative encodings; modest feature selection; transformations	A compact ablation showing what representation changed and whether it helped
C	class weighting; focal-style loss if relevant; sampling / resampling; thresholding logic	Clear evidence of how minority or harder classes changed, even if overall score moved only slightly
D	bagging/boosting variants; calibration checks; soft voting; stacking; robustness checks	A meaningful diagnostic or comparison rather than a large collection of loosely connected trials
E	SHAP / feature importance; subgroup-style fairness checks; error analysis; model interpretation	Concrete insight into model behaviour, not only screenshots

(E) K-Means and Gaussian Mixture Model (GMM) Exploration (6 marks)

This is a compact exploratory section. It is not the main performance section, and it does not require clusters to match the class labels exactly. The aim is to show your understanding of unsupervised learning methods and your ability to interpret their results carefully.

- Use a sensible processed numeric feature space and briefly explain what you clustered on.
- Explore a small range of cluster/component numbers, such as 2-8.
- For **K-Means**, provide sensible supporting evidence, such as inertia (SSE), cluster sizes, or another simple analysis..
- For **Gaussian Mixture Model (GMM),** provide sensible supporting evidence, such as component sizes, posterior confidence/responsibility, or overlap/uncertainty between components.
- Include at least one compact table or figure comparing K-Means and GMM.
- If class labels are used for reference, explain clearly that unsupervised structure does not need to align exactly with supervised labels
- Stronger work may additionally use silhouette score, log-likelihood trends, or a simple visualization.

(F) Final Model Choice and Hidden-Test Export (8 marks)

- Choose the final model using validation evidence only.
- Retrain appropriately using both train and validation dataset and generate the hidden-test CSV in the required format.
- Submit the hidden-test results as **test_result_[your_student_id].csv**. The first column must contain **applicant_id**, the second column must contain **customer_key**, and the third column must contain the predicted **premium_risk** labels (**Standard, High, Low**).
  Incorrect file naming or CSV formatting may prevent automated scoring and will result in an **automatic deduction of 4 marks** from this section.
- Do not tune on the hidden test and do not claim hidden test performance.
- Note: Hidden test score contributes only a small portion of the final marks. High leaderboard rank alone cannot compensate for weak experimental design or poor documentation.

Coursework Answer Sheet / Theory and Reflection (PDF) - all questions below are compulsory (30 Marks)

The Coursework Answer Sheet / Theory and Reflection PDF should not repeat the notebook section by section. All prompt areas below are compulsory. The PDF must be concise, directly linked to your own notebook evidence, and no longer than 4 pages / 1,200 words in total. Exceeding either limit will incur a fixed deduction of 5 marks from the PDF section. You should aim to demonstrate both your theoretical or algorithmic understanding and your experimental findings or practical observations and clearly link your understanding of the algorithms to your experimental analysis. At least one table, figure, or metric from the notebook must be referenced in each theory answer.

Prompt area	What you should do
1. Bagging versus boosting	(1) Briefly state the definitions and key theoretical properties of bagging and boosting models; (2) report the validation results of each model; (3) support your comparison with one or two additional analyses, such as class-wise metrics, a confusion matrix, train–validation behaviour, or stability/sensitivity after tuning; and (4) provide a careful interpretation of what this comparison suggests about this dataset and how it relates to the theoretical properties of bagging versus boosting methods. You are not expected to prove that one model type always performs better.
2. Hyperparameter optimisation	Explain why your optimiser and search space were reasonable for the chosen model, which hyperparameters you expected to matter most, whether the tuned results matched that intuition, and what you learned from the tuning process.
3. K-Means versus Gaussian Mixture Model (GMM)	Explain hard versus soft assignment and the main assumption difference between K-Means and GMM. Then use your own compact evidence to discuss whether the results matched your intuition and whether GMM revealed anything extra, such as soft membership, uncertainty, or a better fit to partial cluster structure.
4. Personalised reflection	Reflect on the compulsory category and on every optional category you implemented. Highlight any unique or interesting algorithm or strategy you tried, the personal challenges you faced, the effort you made to address them, and the key lessons you learned. Honest reflection on a neutral or negative result is acceptable if the reasoning is concrete.

5. AI-use declaration	State briefly what forms of AI assistance, if any, were used. Generic AI-written theory that does not match your notebook evidence will receive limited credit.

Coding Quality, Coursework Answer Sheet Quality, and Submission Guidelines (10 marks)

- Submit your Jupyter Notebook in .ipynb format. It must be well organised, include clear commentary and clean code practices, and show visible outputs. Do not write a second mini-report repeating notebook content.
  - The notebook should be reproducible from start to finish without errors. Results cited in the PDF should be visible in the notebook and should match the reported values.
  - If you used supplementary code outside the notebook, submit that code as well so the full workflow remains reproducible.
- Submit the hidden-test results as test_result_[your_student_id].csv. The first column must contain applicant_id, the second column must contain customer_key, and the third column must contain the predicted premium_risk labels (Standard, High, Low). Incorrect file naming or CSV formatting may prevent automated scoring and will result in an automatic deduction of 4 marks from this section.
- Submit the Coursework Answer Sheet / Theory and Reflection in PDF format. All questions in that section are compulsory. The Coursework Answer Sheet / Theory and Reflection PDF must answer every required prompt, refer to your own notebook evidence, and remain within 4 pages and 1,200 words in total. Exceeding either limit will incur a fixed deduction of 5 marks from the PDF section.
- Include all required components: Jupyter notebooks (code), any additional experimental scripts or custom code, the hidden test-results CSV file, and the Coursework Answer Sheet PDF. Submit all files through the Learning Mall platform. After submission, download your files to verify that they can be opened and viewed correctly to ensure the submission was successful.

Project Material Access Instructions

To access the complete set of materials for this project, please use the links below:

- OneDrive Link: https://1drv.ms/f/c/18f09d1a39585f84/IgCXDMbXkFYSSZUZkkTyXyZzAQ1poX9mujUqF8N3JlL0GD0?e=uNhAHq
- The same coursework materials have also been uploaded to Learning Mall.

When extracting the materials, use the following password to unlock the zip file: **DTS304TC** (case-sensitive, enter in uppercase).