Advanced Machine Learning (MDS)
Evaluation rubric for projects, Fall 2025-26

Lluís Belanche

November 4, 2025

Abstract

This document provides a detailed evaluation rubric for the three independent projects of the Advanced Machine Learning course. The rubric consists of 10 criteria, each evaluated on a 5-level scale. This rubric applies equally to all three projects and emphasizes theoretical rigor, methodological soundness, and clarity of exposition.

1 Grading scale and interpretation

Each criterion is evaluated using five categorical levels with associated numerical ranges:

- **Excellent** (9.0–10.0): Outstanding work that exceeds expectations
- **Very good** (7.0–8.9): Strong work that meets all expectations with minor imperfections
- **Good** (5.0–6.9): Acceptable work that meets basic expectations but has notable limitations
- **Borderline** (3.0–4.9): Work that barely meets minimum standards with significant deficiencies
- **Poor** (0.0–2.9): Work that fails to meet minimum standards
- **Not delivered**[1]: 0

2 Weight intervals

Each criterion has a suggested weight interval (percentage of total project grade). The actual weights may be adjusted by the instructor but should remain within these intervals:

Criterion	**Weight interval (%)**	**25-26**
1. Methodological appropriateness and justification	12–18	15
2. Theoretical rigor and understanding	12–18	15
3. Experimental design and resampling protocol	8–12	10
4. Quality and interpretation of results	10–15	8
5. Critical analysis and comparison of methods	8–12	10
6. Data preprocessing and feature engineering	6–10	8
7. Clarity and organization of written report	8–12	10
8. Scientific conclusions and insights	8–12	10
9. Reproducibility and code quality	6–10	8
10. Use of references and acknowledgment of sources	4–8	6
Total	**100**	**100**

[1] UPC regulations state that the final grade must be NP (not shown) when no evaluation has taken place (exams included).

3 Detailed rubric

3.1 Criterion 1: Methodological appropriateness and justification

Weight: 12–18%

Level	Description
Excellent (9.0–10.0)	Methods chosen are highly appropriate for the problem and course part. Choices are thoroughly justified with clear reasoning connecting problem characteristics, method properties, and theoretical foundations. Demonstrates sophisticated understanding of when and why methods are applicable.
Very good (7.0–8.9)	Methods are appropriate and well-justified. Justifications show good understanding of method properties and problem requirements. Minor gaps in connecting all aspects of the reasoning.
Good (5.0–6.9)	Methods are generally appropriate with basic justification. Some reasoning provided but lacks depth or completeness. May miss some important considerations in method selection.
Borderline (3.0–4.9)	Methods are marginally appropriate or poorly justified. Justifications are superficial or inconsistent. Shows limited understanding of why methods were chosen.
Poor (0.0–2.9)	Methods are inappropriate for the problem or course part, or no justification is provided. Demonstrates fundamental misunderstanding of method applicability.

3.2 Criterion 2: Theoretical rigor and understanding

Weight: 12–18%

Level	Description
Excellent (9.0–10.0)	Demonstrates deep theoretical understanding of all methods used. Mathematical formulations are precise and correct. Assumptions are clearly stated and their implications discussed. Connects theory to practice seamlessly. Provides rigorous analysis of method properties, convergence, or theoretical guarantees where applicable.
Very good (7.0–8.9)	Shows solid theoretical understanding with correct mathematical descriptions. Most assumptions and implications are identified. Minor imprecisions in mathematical notation or theoretical discussion. Good connection between theory and practice.
Good (5.0–6.9)	Demonstrates basic theoretical understanding. Mathematical descriptions are mostly correct but may lack precision. Some assumptions stated but discussion of implications is limited. Adequate but not thorough theoretical treatment.
Borderline (3.0–4.9)	Shows superficial or incomplete theoretical understanding. Mathematical descriptions contain errors or significant imprecisions. Assumptions are unclear or not stated. Weak connection between theory and practice.
Poor (0.0–2.9)	Lacks theoretical understanding. Mathematical formulations are incorrect or absent. No discussion of assumptions or theoretical foundations. Fundamental theoretical errors present.

3.3 Criterion 3: Experimental design and resampling protocol

Weight: 8–12%

Level	Description
Excellent (9.0–10.0)	Experimental design is rigorous and appropriate. Resampling protocol (cross-validation, holdout, etc.) is correctly implemented and well-justified. Statistical significance is properly assessed. Potential sources of bias are identified and addressed. Hyperparameter tuning is done correctly using nested resampling or equivalent.
Very good (7.0–8.9)	Sound experimental design with appropriate resampling protocol. Most aspects of validation are handled correctly. Minor issues in hyperparameter tuning or bias control. Statistical testing is mostly appropriate.
Good (5.0–6.9)	Adequate experimental design with basic resampling protocol. Some aspects may be suboptimal but core methodology is sound. Hyperparameter tuning may have minor flaws. Limited discussion of statistical significance.
Borderline (3.0–4.9)	Experimental design has significant flaws. Resampling protocol is poorly implemented or inappropriate. Hyperparameter tuning is flawed (e.g., using test set). Little attention to statistical validity.
Poor (0.0–2.9)	Experimental design is fundamentally flawed. No proper resampling protocol or serious methodological errors (e.g., data leakage, testing on training data). No consideration of statistical validity.

3.4 Criterion 4: Quality and interpretation of results

Weight: 10–15%

Level	Description
Excellent (9.0–10.0)	Results are presented clearly with appropriate metrics and visualizations. Interpretation is insightful, going beyond surface observations. Uncertainty quantification (confidence intervals, error bars) is provided. Results are contextualized within problem domain and related work. All tables and figures are properly explained.
Very good (7.0–8.9)	Results are well-presented with appropriate metrics. Good interpretation with some insights. Most tables and figures are well-explained. Some uncertainty quantification provided. Minor gaps in contextualization.
Good (5.0–6.9)	Results are presented adequately with basic interpretation. Metrics are appropriate but analysis could be deeper. Tables and figures are present but explanations may be brief. Limited uncertainty quantification.
Borderline (3.0–4.9)	Results are poorly presented or interpretation is superficial. Some tables/figures lack explanation. Metrics may be inappropriate or incomplete. No uncertainty quantification.
Poor (0.0–2.9)	Results are unclear, incomplete, or incorrectly interpreted. Tables and figures are unexplained or misleading. Inappropriate metrics or no proper evaluation.

3.5 Criterion 5: Critical analysis and comparison of methods

Weight: 8–12%

Level	Description
Excellent (9.0–10.0)	Provides thorough, insightful comparison of methods along multiple dimensions (accuracy, interpretability, computational cost, theoretical properties). Critical analysis identifies strengths, weaknesses, and trade-offs. Discusses why certain methods perform better/worse with theoretical or empirical support. Shows mature scientific judgment.
Very good (7.0–8.9)	Good comparison of methods with reasonable critical analysis. Most relevant dimensions considered. Identifies main strengths and weaknesses. Some theoretical or empirical support for observations.
Good (5.0–6.9)	Basic comparison provided focusing mainly on performance metrics. Limited critical analysis. Some strengths and weaknesses mentioned but discussion lacks depth.
Borderline (3.0–4.9)	Superficial comparison with little critical analysis. Focuses only on final numbers without deeper understanding. Misses important trade-offs or differences between methods.
Poor (0.0–2.9)	No meaningful comparison or critical analysis. Methods are not compared, or comparison is fundamentally flawed. No understanding of method differences.

3.6 Criterion 6: Data preprocessing and feature engineering

Weight: 6–10%

Level	**Description**
Excellent (9.0–10.0)	Data preprocessing is thorough, well-justified, and properly executed. All issues (missing values, outliers, scaling, encoding) are appropriately handled with clear reasoning. Feature engineering (if applicable) is creative and theoretically motivated. Preprocessing decisions are clearly documented and justified.
Very good (7.0–8.9)	Good preprocessing covering all major issues. Decisions are justified and mostly appropriate. Feature engineering shows good understanding. Minor issues in execution or justification.
Good (5.0–6.9)	Adequate preprocessing addressing basic issues. Some decisions may be suboptimal but overall approach is reasonable. Limited feature engineering or basic transformations only.
Borderline (3.0–4.9)	Preprocessing is incomplete or poorly executed. Some important issues not addressed. Decisions are poorly justified or inappropriate. Little to no feature engineering when needed.
Poor (0.0–2.9)	Minimal or incorrect preprocessing. Major issues (e.g., data leakage in preprocessing) or important data problems ignored. No justification for decisions.

3.7 Criterion 7: Clarity and organization of written report

Weight: 8–12%

Level	**Description**
Excellent (9.0–10.0)	Report is exceptionally clear, well-organized, and professionally written. Logical flow from introduction to conclusions. Sections are well-balanced. Technical content is explained precisely without unnecessary detail. Writing is concise and free of errors. Figures and tables are well-integrated into text.
Very good (7.0–8.9)	Report is clear and well-organized. Good logical flow with minor structural issues. Writing is generally clear with few errors. Most technical content well-explained. Good integration of figures and tables.
Good (5.0–6.9)	Report is adequately organized but may have structural weaknesses. Writing is understandable but could be clearer or more concise. Some technical content unclear. Figures and tables present but integration could be better.
Borderline (3.0–4.9)	Report organization is weak with poor logical flow. Writing is unclear or verbose. Significant grammar/spelling errors. Technical content difficult to follow. Poor integration of figures and tables.
Poor (0.0–2.9)	Report is poorly organized and difficult to follow. Writing is unclear with numerous errors. Technical content is incomprehensible or missing. Figures and tables are disconnected from text.

3.8 Criterion 8: Scientific conclusions and insights

Weight: 8–12%

Level	Description
Excellent (9.0–10.0)	Conclusions are insightful and well-supported by results. Clearly articulates what was learned and its significance. Identifies limitations honestly and thoughtfully. Suggests meaningful extensions or future work. Reflects deep understanding of the problem and methods. Personal learning is thoughtfully discussed.
Very good (7.0–8.9)	Good conclusions supported by results. Main learnings are clearly stated. Limitations identified. Reasonable suggestions for future work. Shows good understanding and reflection.
Good (5.0–6.9)	Adequate conclusions that summarize main findings. Basic discussion of what was learned. Some limitations mentioned. Limited discussion of extensions or future work.
Borderline (3.0–4.9)	Weak conclusions that barely summarize results. Vague discussion of learning. Limitations not clearly identified or no suggestions for improvements. Little reflection or insight.
Poor (0.0–2.9)	Conclusions missing, trivial, or unsupported by work. No meaningful discussion of learning or limitations. No reflection on the work done.

3.9 Criterion 9: Reproducibility and code quality

Weight: 6–10%

Level	Description
Excellent (9.0–10.0)	Code is well-organized, clearly commented, and follows good programming practices. All results are fully reproducible with clear instructions. Random seeds are set appropriately. Code is efficient and readable. Dependencies are clearly specified. Instructions are complete and easy to follow.
Very good (7.0–8.9)	Code is well-organized and commented. Results are reproducible with minor effort. Good programming practices mostly followed. Instructions are clear and mostly complete. Minor issues with efficiency or documentation.
Good (5.0–6.9)	Code is functional and adequately organized. Results are reproducible but may require some effort. Basic comments provided. Instructions are present but could be clearer. Some code quality issues.
Borderline (3.0–4.9)	Code is poorly organized or documented. Reproducibility is difficult or uncertain. Instructions are incomplete or unclear. Significant code quality issues. Poor programming practices.
Poor (0.0–2.9)	Code is disorganized, undocumented, or non-functional. Results are not reproducible. Instructions are missing or incorrect. Fundamental code quality problems.

3.10 Criterion 10: Use of references and acknowledgment of sources

Weight: 4–8%

Level	Description
Excellent (9.0–10.0)	References are comprehensive, appropriate, and properly cited. All sources (papers, code, data, tools including AI assistants) are acknowledged. Citations are well-integrated into text. Bibliography is properly formatted. Shows engagement with relevant literature. The use of LLMs, if appropriate, is clear and follows recommendations.
Very good (7.0–8.9)	Good use of references with proper citations. Most sources acknowledged. Bibliography mostly well-formatted. Engages with main relevant literature. Minor citation issues. The use of LLMs, if appropriate, follows recommendations with some dubious use.
Good (5.0–6.9)	Adequate references with basic citations. Main sources acknowledged though some may be missing. Bibliography formatting has minor issues. Limited engagement with literature. The use of LLMs, if appropriate, is not totally clear or does not follow recommendations.
Borderline (3.0–4.9)	Insufficient or poorly cited references. Important sources not acknowledged. Bibliography poorly formatted. Minimal engagement with literature. Possible citation issues. The use of LLMs, if appropriate, is insufficient or wrong.
Poor (0.0–2.9)	References missing, inappropriate, or improperly cited. No acknowledgment of sources. No engagement with literature. Serious citation problems or potential plagiarism. The use of LLMs, if appropriate, is absent (this is considered plagiarism and cheating).

4 Notes on application

- This rubric applies equally to all three projects of the course.
- For projects with minimal or no empirical work (more theoretical/analytical focus), criterion 3 (experimental design and resampling protocol) should be interpreted as rigor of theoretical/analytical methodology, and criterion 9 (reproducibility and code quality) may have reduced weight, with corresponding increases in criteria 2 (theoretical rigor) and 8 (scientific conclusions).
- Students should be aware that theoretical rigor (criterion 2) is heavily weighted and applies to all types of work, whether empirical or analytical.
- The instructor may adjust weights within specified intervals based on the specific nature and requirements of each project.
- Intermediate scores within each level range (e.g., 7.5 for very good) can be considered the most probable grade, but can be altered to reflect nuanced performance.