Explainable AI with Shapley value - Semester project

Project title: Master's Thesis Proposal + Realization Plan

Points: 16

Goal

The goal of this project is to practice the early stages of a Master's thesis: topic selection, information elicitation, problem formulation, research planning, and structured academic writing.

Your topic must be centered on AI/ML models and must include a substantial explainability component based on Shapley value (SHAP). No coding is required and not expected for this course project.

You are graded on the quality of your proposal and your ability to justify design choices.

1) Topic requirements

1.1 AI/ML core

The thesis proposal must involve a clearly defined task (classification, regression, forecasting, planning).

1.2 Shapley-based explainability component

Your proposal must include Shapley value as an explainability method. It must address:

- What is being explained (a single decision, a model globally, etc.)
- Who the explanation is for (developer, deployer/user, affected person, regulator, domain expert)
- What Shapley value(s) you will produce (e.g., local feature attributions, global summaries)
- Known limitations (baseline/background choice, feature dependence/correlation, stability)

1.3 Feasibility

The proposal must be feasible as a master's thesis (4–6 months of full-time work or equivalent). You must identify practical constraints such as data access, compute, and evaluation setup.

2) Deliverable (what to submit)

Submit one PDF containing three parts (in this order). Total length max 4 pages (including figures/tables; excluding optional cover page).

Part A - Thesis proposal (max 1 page)

Mandatory structure (use these headings):

1. Title + at most 5 keywords
2. Motivation and context
3. Problem statement
4. Proposed approach

5. Shapley explainability component (must be specific)
6. Expected contributions
7. Key references

For the “Shapley explainability component” item, you must include:

- What Shapley/SHAP variant or viewpoint you will use
- What players/features are, and what output you explain

Part B - Detailed realization plan (suggested 1–2 pages)

Mandatory content:

- Activities (for example, literature reading, problem formalization, baseline model, SHAP pipeline, evaluation, writing)
- Timeline
- Evaluation plan: datasets (or how you will obtain data), metrics, baselines; explanation evaluation
- Risks and mitigations (data access risk, feature dependence, unstable attributions, data availability)
- Information elicitation plan: Identify at least 2 people/groups you could consult (e.g., potential supervisor, PhD student, domain expert, industry contact).

Part C - Thesis content outline (max 1 page)

Provide a proposed table of contents with 8–12 sections. For each section, add one sentence describing what it will contain.

Minimum required chapters/sections:

- Introduction + problem formulation
- Background on Shapley/SHAP (and related explainability methods you compare to)
- Proposed method / approach
- Experimental setup + evaluation methodology
- Results + discussion (including limitations)
- Conclusion + future work

3) Sources and writing rules

3.1 Sources

- Minimum 6 references, maximum 10 in Part A.
- Each research question/problem must be motivated by at least one cited source.

3.2 Writing style

Write like a thesis: clear claims, supported by evidence, no marketing language.

Avoid vague words (“improve”, “better”, “robust”) without defining what metric and what baseline.

4) LLM policy

LLM tools (ChatGPT, Claude, etc.) may be used, but you must include an LLM-use disclosure at the end of the PDF:

- Tool(s) used
- What you used them for (outline, grammar, brainstorming, etc.)
- What you verified manually (facts, citations, technical claims)

Important: You must be able to explain and defend your proposal in your own words.

5) Short defence

Each student will have a brief individual discussion with the instructor (in person or online):

- Your pitch (what + why + what is new + where Shapley fits)
- Questions (Shapley semantics, feasibility, evaluation plan, sources)

This is primarily for feedback and to practice thesis-style defence.

6) Grading (16 points total)

Written PDF (12 points)

- (4 pts) Thesis proposal quality (Part A)
- (4 pts) Shapley explainability integration
- (2 pts) Realization plan (Part B)
- (2 pts) Thesis outline (Part C)

Short defence (4 points)

Ability to justify choices and the content of the proposal

References

1. J. Pecka. Cooperative game theory for machine learning task*s, Master Thesis*, 2020, http://hdl.handle.net/10467/98471
2. Covert, I., Lundberg, S., & Lee, S. I. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, *22*(209), 1-90, 2021. https://www.jmlr.org/papers/volume22/20-1316/20-1316.pdf
3. https://shap.readthedocs.io/en/latest/#