Gemini End-to-End Machine Learning Evaluation Report

1. Objective

The objective of this project is to evaluate whether Google Gemini can reliably execute a complete end-to-end machine learning workflow when provided with a strict task prompt. The focus is on instruction adherence, reproducibility, and artifact generation rather than model accuracy.

2. Updated Project Instructions

The project follows the updated instructions provided by the evaluator: a prompt of 300–600 words was designed first, followed by unit tests derived strictly from the prompt. A simple reference solution was implemented locally, and Gemini was evaluated using the same prompt and tests in Google Colab.

3. Prompt-First Methodology

The file prompt.md serves as the single source of truth. All requirements regarding data preprocessing, feature engineering, model training, evaluation, and output artifacts are explicitly stated in the prompt. Unit tests were written only after the prompt was finalized.

4. Unit Testing Strategy

Five pytest-based unit tests were created to verify the existence and correctness of the required output artifacts. These tests were intentionally minimal and directly mapped to the prompt requirements, ensuring objective and reproducible evaluation.

5. Reference Solution

A correct local reference solution was implemented to confirm that all prompt requirements are achievable. The reference solution successfully passed all unit tests, validating the correctness of the tests and the clarity of the prompt.

6. Gemini Evaluation (Colab)

The same prompt was provided to Gemini in a Google Colab environment without any human intervention or clarification. Gemini performed exploratory analysis and trained a Logistic Regression model but failed to generate the required artifacts with the specified names and formats.

7. Test Results on Gemini Output

When the unit tests were executed against Gemini’s output, all tests failed due to missing or incorrectly named files. This demonstrates Gemini’s inability to strictly follow end-to-end machine learning instructions.

8. Submission Contents

- experiment/: Contains prompt.md, test_notebook.py, requirements.txt, and the dataset.
- colab_notebook.ipynb: Google Colab notebook used for Gemini evaluation.
- screenshots/: Evidence of Gemini outputs and pytest failures.
- README.md: Brief summary of the experiment and methodology.

9. Conclusion

This experiment demonstrates that while Gemini can perform individual machine learning tasks, it does not reliably comply with strict procedural and artifact-generation requirements. Automated unit testing is therefore essential when evaluating LLM-driven data science workflows.