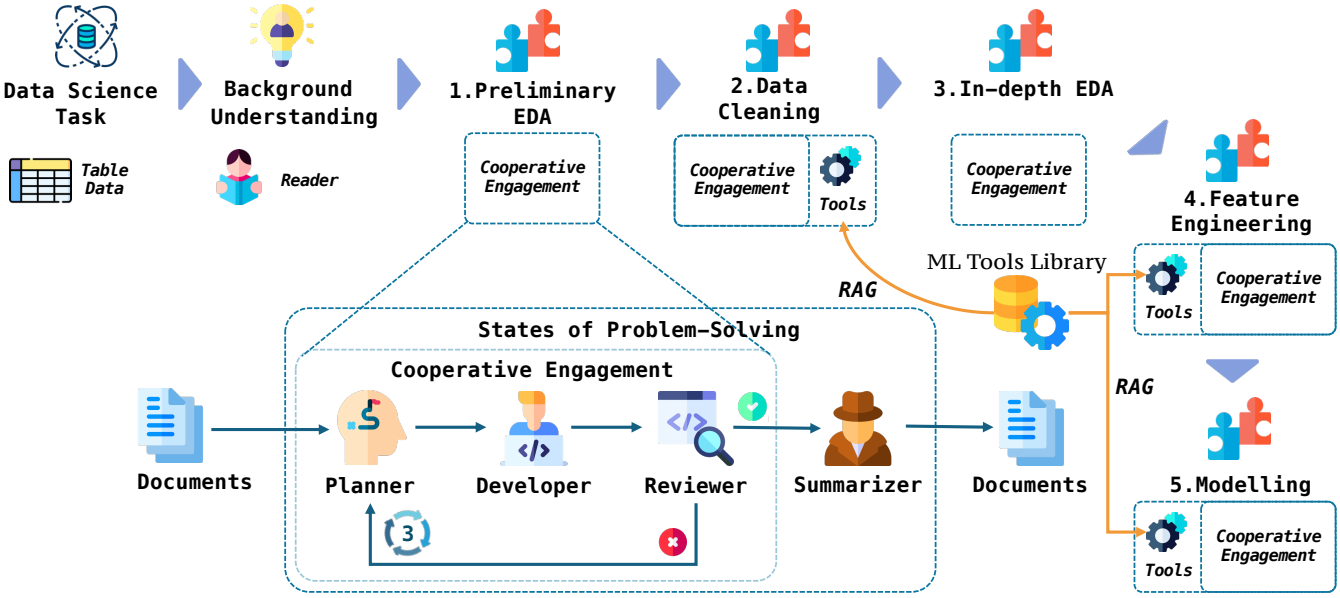
AutoKaggle: A Multi-Agent Framework for Autonomous Data Science Competitions

Data Science

What: Data science is an interdisciplinary field that extracts valuable information, insights, and supports decision from data.
How: process of data science from *understanding the background*, *preliminary exploratory data analysis*, *data cleaning*, *in-depth exploratory data analysis*, *feature engineering* to *model-building, -validation, and –prediction.*

AutoKaggle



AutoKaggle provides a universal and comprehensive solution for various data science tasks in tabular formats.

Features:

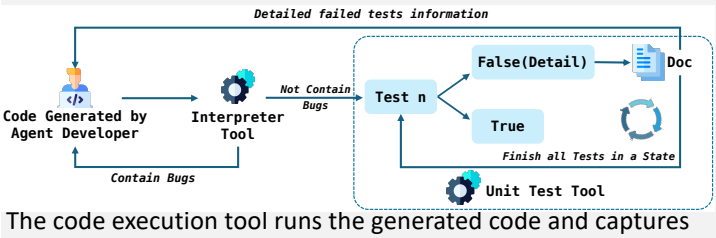
- Phase-based Workflow and Multi-agent Collaboration.
- Iterative Debugging and Unit Testing.
- Machine Learning Tools Library.
- Comprehensive Reporting.

Machine Learning Tools

Generating machine learning code from scratch using LLMs can be challenging due to the intricacies of various tasks.

Our machine learning library reduces the burden in detailed programming tasks, enabling to focus more on higher-level task planning and code design.

Debugging and Unit Testing



The code execution tool runs the generated code and captures any runtime errors.

Debugging: analyze error messages and attempts to fix the code.

- Error localization
- Error correction
- Corrected code segments
- Merging segments to the old codes

Unit Testing: run predefined tests to ensure the code meets requirements.

Evaluation

Datasets: select eight Kaggle competitions on classification and regression tasks.

Metrics:

- **Made Submission (MS):** percentage of times a submission file generated.
- **Valid Submission (VS):** percentage of submission files can be successfully submitted to the Kaggle website.
- **Normalized Performance Score (NPS):** normalization for bounded metrics and unbounded metrics.

$$\text{NPS} = \begin{cases} \frac{1}{1+s}, & \text{if } s \text{ is smaller the better} \\ s, & \text{otherwise.} \end{cases}$$

- **Comprehensive Score (CS):** evaluate both the pass rate and the average performance rate.

$$\text{CS} = 0.5 \times \text{VS} + 0.5 \times \text{ANPS}$$

Main Results:

Metric	Setting / Task	Classic				Recent				
		Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Avg.
Made Submission	AutoKaggle gpt-4o	1	0.80	0.80	1	0.80	0.80	0.80	0.80	**0.85**
	AutoKaggle o1-mini	1	0.60	0.60	1	0.60	0.80	0.60	0.60	0.73
	AIDE gpt-4o	1	0.40	0.20	0.60	1	0.80	0.80	0	0.60
Valid Submission	AutoKaggle gpt-4o	1	0.80	0.80	1	0.80	0.60	0.80	0.80	**0.83**
	AutoKaggle o1-mini	1	0.60	0.60	1	0.60	0.60	0.60	0.60	0.70
	AIDE gpt-4o	1	0.40	0.20	0.40	1	0.80	0.80	0	0.58
Comprehensive Score	AutoKaggle gpt-4o	0.888	0.786	0.831	0.862	0.810	0.728	0.848	0.812	**0.821**
	AutoKaggle o1-mini	0.879	0.680	0.729	0.863	0.709	0.735	0.742	0.735	0.759
	AIDE gpt-4o	0.872	0.597	0.542	0.561	0.918	0.793	0.848	0	0.641

AutoKaggle (GPT-4o) demonstrated superior performance, surpassing the AIDE framework by 28%. In AutoKaggle, the GPT-4o achieved better results than the o1-mini.

Codes: **https://m-a-p.ai/AutoKaggle.github.io/**