Practical Machine Learning

プロジェクト概要 -Project Overview-

本プロジェクトの目的は、Kaggleのコンペティションに取り組むことで、
各コンペティションの目標の達成に貢献するとともに、機械学習のスキルを習得することである。
後期には、2つのグループに分かれて、それぞれ別のコンペティションに挑戦した。

The goal of this project is to take part in Kaggle competitions, work on their tasks, and develop our machine learning (ML) skills. In the second semester, we split into two groups, and each group worked on a different competition.

-Member-

Motohide Wada　Rikuto Mori
Yuki Kon　Yuto Sasaki
Kenta Yoshizaki　Haruto Sato
Ayane Onishi　Shun Sasaki

-Adviser-

Hirokazu Shimauchi
Masaaki Shirase
Hiroshi Yamada
Satoshi Kawaguchi
Naoyuki Sato

Kaggleとは -What is Kaggle-

機械学習コンペティションのプラットフォームである。企業や研究機関が出題したデータを使い、予測モデルの精度を競い合うことで実践的なスキルを身につけられる。

Kaggle is a platform for ML competitions. By using data provided by companies and research institutions to compete for higher prediction accuracy, participants can gain practical ML skills.

Thermophysical Property: Melting Point -Team A-

コンペティションの目的 -Purpose-

分子記述子を用いて、有機化合物の融点を予測する機械学習モデルの構築を目的とする。有機化合物の融点予測は、薬剤設計や材料探索、プロセス安全評価などの応用へつながる知見を得ることが期待できる。

The goal of this project is to build an ML model to predict the melting points of organic compounds using molecular descriptors. Such predictions are expected to provide insights that can contribute to applications in drug design, materials discovery, and process safety evaluation.

説明変数

①SMILES：分子内の原子の並びや結合関係を文字列で表現したもの
②Group：分子の種類や構造的特徴をまとめるために付けた分類ラベル

目的変数

③Tm：融点

Explanatory variables

①SMILES: a string that represents the atoms and bonds in a molecule
②Group: a label used to summarize the type or structural features of the molecule

Target variable

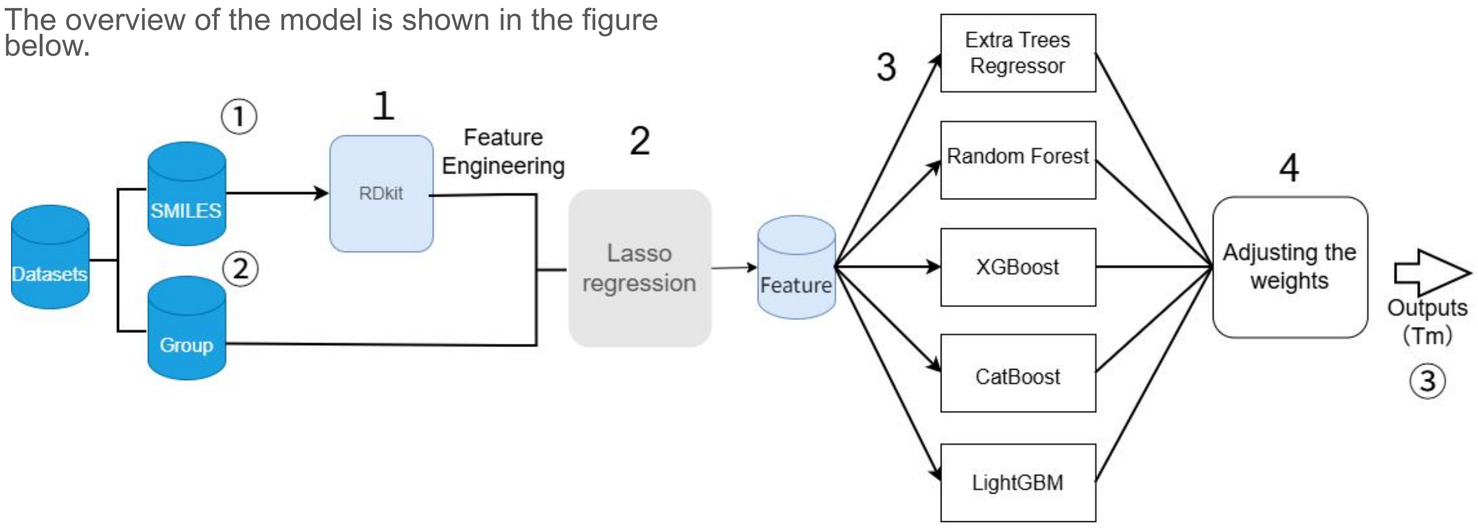
③Tm: melting point

手法 -Approach-

1. RDkitを用い特徴量エンジニアリングを行う。
2. Lasso回帰を用い特徴量を選定する。
3. 選定された特徴量を用いて複数の決定系のモデルを構築する。
4.各モデルの重みを調整しアンサンブル学習を行う。
モデルの概要は下図の通りである。

1. Feature engineering was performed using RDKit
2. Features were selected using Lasso regression.
3. Multiple tree-based models were built using the selected features.
4. The weights of each model were adjusted, and ensemble learning was performed.

The overview of the model is shown in the figure below.



結果 -Result-

MAE:23.52665

2025/11/21現在の上位10%以上の精度を持つモデルを構築できた。決定木系のモデルを使用したこととLasso回帰を用いたことが、今回のモデルの精度向上に寄与していると思われる。畳み込みニューラルネットワークの実装や、AutoGluonを使用しより最適なパラメータの探索を行うことで、スコアを向上させることができる可能性がある。

As of 2025/11/21, we built a model with a score in the top 10%. The use of tree-based models and Lasso regression likely contributed to this performance. Implementing convolutional neural networks or using AutoGluon to optimize parameters may further improve the score.

ARC Prize 2025 -Team B-

コンペティションの目的 -Purpose-

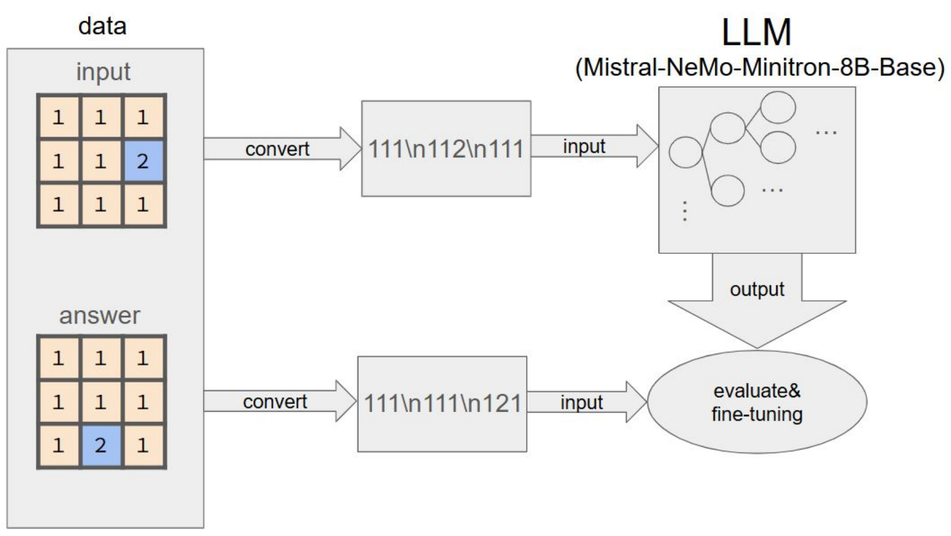
ARCテストとは、汎用人工知能のベンチマークテストとして考案されたものである。現在のAIは、少ないデータから新しい法則を見つけることは難しく、ARCテストではほとんど得点できていない。このコンペティションでは、そのテストを解くモデルの構築を目指す。

ARC tests were devised as benchmark tests for artificial general intelligence. It is difficult for current AI models to discover new rules from limited data, and they barely score on the ARC test. This competition aims to build a model that can solve the ARC test.

手法 -Approach-

1.入力する行列データを文字列に変換する。
2.オープンソースのLLMである
Mistral-NeMo-Minitron-8B-Baseを基盤に
ARCテストで事前学習したモデルを採用する。
3.そのモデルについて、変換したデータを用いて
ファインチューニングを行う。
手法の概要は下図の通りである。

1. Convert the input matrix data into a string.
2. Adopt a model pre-trained on the ARC test based on Mistral-NeMo-Minitron-8B-Base, one of the LLMs.
3. Fine-tune that model using the transformed data.
The overview of the Approach is shown in the figure below.



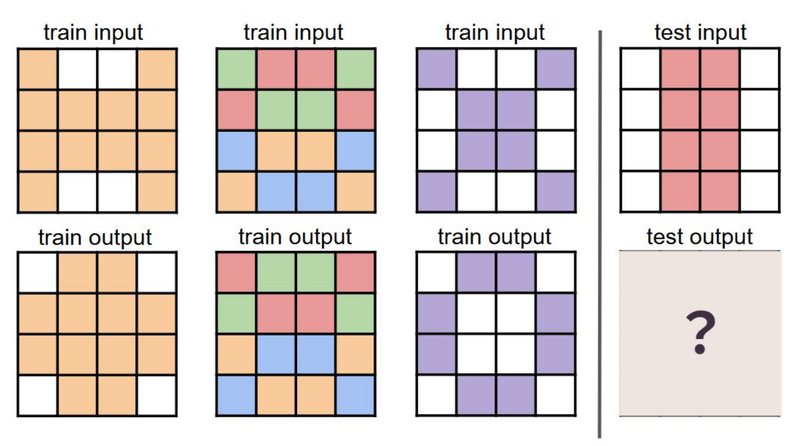
結果 -Result-

スコア：5.83/100
上位10%

Score: 5.83/100
Top 10%

手法としてLLMを用いることにより、抽象的なアルゴリズムを学習することが出来た。また、計算時間に制約があったため、GPUのタスクの量を均等にした。その結果、処理を効率よく行うことが出来た。このようなモデルを構築することにより、上位10%に入ることが出来た。

By using an LLM as the method, we were able to learn abstract algorithms. Also, because of the computational time limitation, we evened out the number of tasks on the GPU. As a result, we were able to process the tasks efficiently. By using this model, we were able to rank in the top 10%.



問題の一例　An example of a task