Practical Machine Learning

プロジェクトリーダー：和田 基秀/WADA Motohide

1　背景

近年、人工知能（AI）は急速に発展しており、医療や製造業、教育など幅広い分野で活用が進んでいる。中でも機械学習は、計算機科学を基盤とする。AI 技術の中核を担う重要な要素である [1]。データに基づいて課題を分析し、仮説の検証と改善を行う能力は今後ますます重要となる。機械学習は、理論と実践を結び付けて学ぶことが不可欠な技術である。

そこで本プロジェクトでは、Kaggle 上のコンペティションに参加し、よりよいモデルを構築することでコンペティション自体の目標の達成に貢献することを目的として活動を行った。本活動を通じて、機械学習について知識を深め、実践的なスキルの獲得したことに加え、チームで協働しながら問題解決に取り組む力の育成を目指した。

2　関連技術

2.1　機械学習

機械学習とは、データから規則性を学習し、それに基づいて予測や分類を行う技術であり、画像認識や自然言語処理など多くの分野で利用される AI の中核的技術である [1]。

機械学習は、教師あり学習、教師なし学習、強化学習に大別され、目的やデータの性質に応じて適切な手法が選択する。線形回帰や決定木、ニューラルネットワークなど多様な手法が存在し、過学習を防ぎ汎化性能を高めるために、正則化や交差検証によりモデルを調整することが重要となる。

2.2　Kaggle

Kaggle は、機械学習を用いて課題に取り組むコンペティションが開催されているプラットフォームである。企業や研究機関が提供する実データを用いたコンペティションを通じて、モデル構築から評価までの一連のプロセスを実践的に経験できる点が特徴である。また、他の参加者の手法や結果を参考にしながら学習を進められるため、機械学習の理解を深める上で有用な環境である。

3　活動

3.1　読み会・入門的なコンペティション

まず基礎的な知識を身につけるために、『Python で始める Kaggle スタートブック』[2] を用い、読み会形式で学習を行った。書籍を通して、分類および回帰といった機械学習の主要手法や、特徴量作成、モデル評価、交差検証の重要性について理解を深めた。特に、過学習を防ぎ汎化性能を高めるための正則化やモデル選択の考え方を、実践的に学ぶことができた。

その後、実践的な機械学習の基礎習得を目的として、2 つのチームに分かれ、自然言語処理などを題材とした 4 つのコンペティションに参加した。

3.2　前期の本番コンペティション

前期の後半では、前半で学んだ内容を生かして、2 つのチームに分かれて引き続き機械学習を学びながら、企業が提供するコンペティションに取り組んだ。参加した 2 つのコンペティションは、暗号資産の市場データから、次の 1 分間における価格変動を予測する DRW -Crypto Market Prediction-[3] と、手首装着型センサーから取得された時系列データから、身体集中反復行動 [4] と日常的なジェスチャーを分類する CMI -Detect Behavior with Sensor Data-[5] である。

3.3　後期の活動

後期においても 2 つのチームに分かれてコンペティションに参加した。1 つ目のチームは Thermophysical Property: Melting Point に、2 つ目のチームは ARC（Abstraction and Reasoning Corpus）Prize 2025 に取り組んだ。4 章及び 5 章において両コンペティションに対する取り組みを記述する。

4　Thermophysical Property : Melting Point

4.1　コンペティションの概要・背景

本活動では、Kaggle 上で公開されている Thermophysical Property: Melting Point コンペティション

に取り組み、有機化合物の分子記述子を用いた融点予測モデルの構築を行った。融点は薬剤設計や材料選定、化学プロセスの安全性評価において重要な物性値であるが、実験的測定には多大な時間とコストを要する [6]。そのため、分子構造情報を基に機械学習によって融点を予測する手法が近年注目されている。

4.2　目的

本コンペティションの目的は、学習データから機械学習モデルを構築し、未知の有機化合物に対しても高精度に融点を予測可能な手法を検討することである。学習データに含まれる説明変数として、有機化合物の構造式を線形文字列で表現した SMILES（Simplified Molecular Input Line Entry System）、官能基数や基本的な分子記述子を表す Group、各有機化合物に付与された識別子である ID が与えられ、目的変数は有機化合物の融点 Tm（ケルビン）である。評価指標には平均絶対誤差（Mean Absolute Errorh: MAE）が採用されている。本活動を通じて、分子記述子の扱い方や特徴量エンジニアリング、回帰モデル構築および評価手法の理解を深めることを目指した。

4.3　手法

1. RDKit を用いた特徴量エンジニアリング

SMILES で表現された分子構造を RDKit[7] により分子オブジェクトへ変換し、機械学習で扱える数値特徴量を生成した。分子オブジェクトには原子情報、結合関係、環構造などが保持されており、これを基に特徴量計算を行った。

分子記述子としては、RDKit が標準で提供する 208 種類すべてを使用し、分子量、電子状態、極性、環構造、分子形状、官能基フラグメントの出現回数など、分子の物理化学的および構造的特徴を数値化した。さらに、官能基や構造モチーフの有無を表す MACCS Keys（Molecular ACCess System Keys）を 167 次元と、原子周辺の局所構造を表現する Morgan フィンガープリントを半径 2、3、4 の 3 条件で算出し、それぞれ 512 次元の 1,536 次元を導入した。

これらを統合することで、各分子から合計 1,911 次元の特徴量を構築した。

2. LASSO 回帰による特徴量選択

前節で構築した 1,911 次元の特徴量は高次元であり、冗長な情報を含む可能性がある。そこで、特徴量選択を目的として RFE（Recursive Feature Elimination）、KBest（SelectKBest）、LASSO 回帰（Least Absolute Shrinkage and Selection Operator）を用いて特徴量選択を行った。結果として LASSO 回帰の精度が最も高かったため、LASSO 回帰を用いて最終的なモデルを構築した。LASSO 回帰は、回帰係数に L1 正則化を課すことで一部の係数を 0 とし、高次元データにおける変数選択を可能とする手法である。すべての特徴量を標準化した上で、5 分割交差検証により正則化強度を自動決定する LassoCV を適用し、回帰係数が 0 でない特徴量のみを選択した。その結果、特徴量数は 1,911 次元から 347 次元へ削減され、融点予測に有効な分子サイズ、極性、官能基構成、局所構造に関する情報が保持された。

その結果、特徴量数は 1,911 次元から 347 次元へ削減され、融点予測に有効な分子サイズ、極性、官能基構成、局所構造に関する情報が保持された。

3. 回帰モデルとアンサンブル

選定された特徴量を用いて，Random Forest，Extra Trees，XGBoost，LightGBM，CatBoost の 5 種類の決定木系モデルを構築した。その概観を図 1 に示す。さらに、各モデルの予測結果を重み付きで統合するアンサンブル学習を行い、単体モデルを上回る性能向上を確認した。表 1 がその結果である。

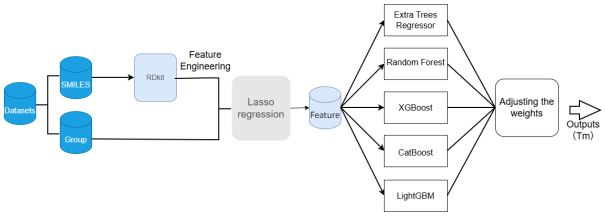

図 1: モデルの概観図

表 1: モデル性能の比較（MAE）

Model	MAE
Extra-Trees	27.8162
RandomForest	29.9569
XGBoost	26.8549
CatBoost	27.0117
LightGBM	27.8270
Ensemble (Best)	25.7793

4.4　結果と考察

最終的に MAE 23.53 を達成し、2026 年 1 月 7 日現在上位 10% 以上に相当する精度のモデルとなった。RDKit による分子特徴量生成と LASSO 回帰による特徴量選択、ならびに決定木系モデルのアンサンブルが精度向上に寄与したと考えられる。深層学習モデ

ルをモデルに導入することや AutoML（Automated Machine Learning) 手法を用いてモデル選択やモデルのパラメーター調整を行うことにより、より精度の高いモデルとなることが期待できる。

5 ARC Prize 2025

5.1 背景

本活動の主題である汎用人工知能.（Artificial General Intelligence：AGI）とは、人間と同等、あるいはそれ以上に、広範な領域において自律的に学習・理解し、問題を解決できる知能を指す [8]。従来の AI が特定のタスクに特化して発展してきたのに対し、AGI は未知の状況に対しても柔軟に適応できる能力が特徴である。

一方で、近年の大規模言語モデル.（Large Language Model：LLM）をはじめとする AI 技術の実態は「膨大な学習データに基づく統計的なパターンマッチング [9]」に依存している側面が強い。そのため抽象化能力や論理的推論能力においては、依然として大きな課題が残されている。

このような能力を測定するベンチマークとして考案されたのが ARC テストである。これは、単純な視覚的ルールを、数個の例示から推論して正解を導き出すことを求めるものである。ところが、最新の AI モデルであっても人間並みの正解率を達成することは極めて困難であり、多くの既存モデルが 20 %以下に留まっている [10]。

5.2 目的

そこで本プロジェクトでは、ARC Prize 2025 への参加を通じ、現在の AI が直面している推論能力の限界を打破することを目指した。

5.3 コンペティションの概要

本コンペティションでは、ARC テストの正解率を競う。各タスクは、数組の入力画像（Input）と出力グリッド（Output）のペアで構成されており、参加者は提示されたペアから変換ルールを推論し、テスト用の入力グリッドに対する正確な出力を予測するモデルを構築する。

ここで扱われるデータは、0 から 9 までの整数で構成される二次元行列であり、そのサイズは最小 1 × 1 から最大 30 × 30 までと可変的である。表 2 が与えられるデータである。

また、本コンペティションは実行を 12 時間以内に完了させなければならないという制約がある。

表 2: データ形式

種類	タスク数	正解ラベル
学習用データ（Train）	1000	有
評価用データ（Test）	120	無

5.4 手法

1． 特化型ソルバーの構築

本手法では、ARC-AGI-2 の学習用データにおける解法を体系化し、導出した解法パターンによるタスク分類を試みた。まず、学習用データを実際に人力で解き、解法で重要なキーワードを特徴量として抽出した。続いて、この特徴量空間上で K-Means 法によるクラスタリングを行った。しかし、結果を定性的に評価したところ、複数のクラスタにおいてどの定義済み特徴量にも当てはまらない、もしくは複数の解法が複合するといった抽象的なタスクが多く存在することが判明した。そのため、特定できた特徴量に対応する特化型ソルバーを作成し、該当するクラスタ内のタスクに対して適用することで、部分的な正解率の向上を図った。

2． LLM を用いた推論と計算資源の最適化

本手法では、ARC Prize 2024 の上位入賞者の手法を参考に、LLM を活用した解法を採用した。

まず、視覚的なグリッドデータを LLM で処理可能にするため、二次元行列を改行文字を含む一連の文字列へと変換する前処理を行った。これにより、自然言語処理における系列変換タスクとして再定義した。

モデル構築においては、自然言語やプログラムコードを含む大規模なデータセットで事前学習済みの LLM である Mistral-NeMo-Minitron-8B-Base[11] をベースとし、提供された 1,000 問の学習データを用いてファインチューニングを実施した。図 3 に構築したモデルの概要を示す。

推論フェーズでは、生成された予測結果を行列形式に再変換し解答とした。その際、大会規定である 12 時間の制限を最大限に活用するため、並列処理の最適化を行った。複数の GPU 間で演算負荷が均等になるよう処理を動的に配分することで、GPU の待機時間を最小限に抑え、時間内での試行回数と処理速度を最大化させた。

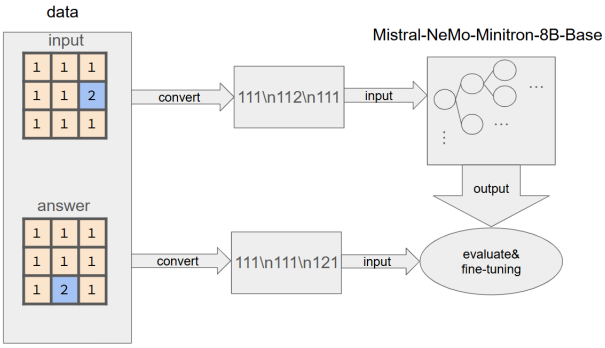

図 2: モデルの全体図

5.5 結果

特化型ソルバーは学習用データで正解率 13.6 %、評価用データで 0 %であった。一方、ファインチューニングした LLM は評価用データで 5.83 %を達成した。

5.6 考察

特化型ソルバーの学習用データと評価用データでのスコアの乖離は、開発したソルバー群は特定のルールセットには有効であったが、未知の法則性が求められる初見の問題に対しては適応力を欠いていたことが原因であると考えられる。この結果より、ルールベースに近い特化型のアプローチは、AGI の本質である未知の課題への適応能力においては限界がある可能性が示唆される。

特化型アプローチとは対照的に、LLM を用いたアプローチでは、事前学習された LLM を学習用データに合わせてファインチューニングするアプローチが有効に機能した。この結果は、視覚的なグリッド情報を言語テキストとして処理させる手法が、未知の抽象的なルールを推論する上で有効に機能することを示している。また、技術的な側面として、GPU の並列処理における負荷分散の最適化が結果に大きく貢献した。計算リソースを最大限に活用し、12 時間という制限時間内で推論試行回数を最大化できたことが結果につながった。

6 総括

本プロジェクトでは、Kaggle のコンペティション自体の持つ目標に貢献することを通じて、機械学習の実践的スキルの獲得と、チームによる問題解決能力の育成に取り組んだ 。特に後期に取り組んだコンペティションにおいては、それぞれのコンペティションで上位 10 %以上の精度を持つモデルを構築できた。客観的な成果指標において一定の評価を得た。

参考文献

[1] C.M.Bishop 著：『パターン認識と機械学習 上 ベイズ理論による統計的予測』，元田浩，栗田多喜夫，樋口知之，松本裕治訳，講談社，2020.

[2] 石原祥太郎，村田秀樹：『Python で始める Kaggle スタートブック』，講談社，2020.

[3] Kaggle: “DRW Crypto Market Prediction”, https://www.kaggle.com/c/drw-crypto-market-prediction（参照日：2026 年 1 月 14 日）.

[4] 境玲子，飯田美紀：「皮膚・毛髪への“身体集中反復行動”—抜毛症，皮膚むしり症，皮膚の掻破行動—」,『児童青年精神医学とその近接領域』，Vol. 57, No. 2, pp. 298–309, 2016.

[5] Kaggle: “Child Mind Institute - Detect Behavior with Sensor Data”, https://www.kaggle.com/competitions/cmi-detect-behavior-with-sensor-data（参照日：2026 年 1 月 14 日）.

[6] Kaggle: “Thermophysical Property: Melting Point”, https://www.kaggle.com/competitions/melting-point/overview（参照日：2026 年 1 月 9 日）.

[7] RDKit Open-Source Cheminformatics Software: “An Overview of the RDKit”, https://www.rdkit.org/docs/Overview.html（参照日：2025 年 12 月 17 日）.

[8] ソフトバンク：「汎用人工知能（AGI）とは？特化型 AI との違いや開発状況、リスクについて解説」，SoftBank Business Blog, 2023-10-24, https://www.softbank.jp/business/content/blog/202310/what-is-agi（参照日：2026 年 1 月 7 日）.

[9] F.Chollet: “On the Measure of Intelligence”, arXiv:1911.01547 [cs.AI], 2019.

[10] ARC Prize: “ARC-AGI-2”, https://arcprize.org/arc-agi/2/（参照日：2026 年 1 月 7 日）.

[11] NVIDIA: “Mistral-NeMo-Minitron-8B-Base Model Card”, NVIDIA API Catalog, https://build.nvidia.com/nvidia/mistral-nemo-minitron-8b-base/modelcard（参照日：2026 年 1 月 7 日）.