公立はこだて未来大学 2025 年度 システム情報科学実習
グループ報告書

Future University Hakodate 2025 Systems Information Science Practice
Group Report

プロジェクト番号/Project No.

6

プロジェクト名

Practical Machine Learning

Project Name

Practical Machine Learning

プロジェクトリーダー/Project Leader

和田基秀　WADA Motohide

グループリーダー/Group Leader

森陸飛　MORI Rikuto

佐藤大翔　SATO Haruto

グループメンバー/Group Member

佐々木悠光　SASAKI Yuto

大西彩音　ONISHI Ayane

今優稀　KON Yuki

良崎健太　YOSHIZAKI Kenta

佐々木舜　SASAKI Shun

指導教員

島内宏和　山田浩　白勢政明　川口聡　佐藤直行

Advisor

SHIMAUCHI Hirokazu　YAMADA Hiroshi　SHIRASE Masaaki　KAWAGUCHI Satoshi

SATO Naoyuki

提出日

2026 年 1 月 21 日

Date of Submission

Jan. 21, 2026

概要

本プロジェクトの目的は、機械学習のコンペティションに参加し、実践的な課題にチームで取り組むことを通じて、機械学習に関する理論的知識および実装技術を体系的に習得するとともに、コンペティションにおいて有意義な成果を上げることで、各コンペティションの目標に貢献することである。前期においては、参考書を用いた勉強会や入門レベルのコンペティションへの参加を通じて、機械学習の基礎的知識とモデル構築の一連の流れを習得した。また、2 つのチームに分かれ、企業が提供する実践的なコンペティションに挑戦することで、目的に応じた特徴量設計およびモデル選択の重要性などについて理解を深めた。後期においては、前期の活動を通じて得られた知識および経験を活用し、2 つのチームに分かれ、実践的なコンペティションに本番として取り組んだ。1 つ目のグループでは、有機化合物の融点予測に関するコンペティションに取り組み、薬剤設計や材料探索、プロセス安全評価などへの応用につながりうる予測モデルの構築を目指した。2 つ目のグループでは、汎用人工知能のベンチマークテストを解くモデルの構築に取り組んだ。両グループとも、コンペティション内において、上位 10 % 以上の精度を持つモデルを構築することができた。

キーワード　機械学習, Kaggle, 探索的データ分析, 化学情報学, 汎用人工知能

Abstract

The objective of this project is to systematically acquire theoretical knowledge and implementation skills in machine learning through participation in competitive challenges as a team. Furthermore, the project aims to contribute to the goals of each competition by achieving significant results.In the first half of the year, we acquired fundamental knowledge and learned the standard workflow of model development through study groups using textbooks and participation in introductory competitions. Additionally, the project members were divided into two teams to tackle practical competitions provided by corporations, deepening our understanding of the importance of feature engineering and model selection tailored to specific objectives.In the second half of the year, leveraging the knowledge and experience gained during the first half, the two teams engaged in practical competitions as their primary focus. The first group participated in a competition to predict the melting points of organic compounds, aiming to build predictive models that could be applied to drug design, material discovery, and process safety assessment. The second group focused on developing a model to solve benchmark tests for Artificial General Intelligence (AGI).Both groups were able to build models with accuracy in the top 10 % or higher within the competition.

Keyword Machine learning, Kaggle, Exploratory Data Analysis, Chemoinformatics, Artificial General Intelligence

目次

第 1 章　はじめに

1.1　背景

近年、人工知能 (AI:Artificial Intelligence) は急速に発展しており、医療、金融、マーケティング、製造業、教育など、社会のあらゆる分野で実用化が進んでいる。特に計算機科学の分野を起源とする機械学習は、AI 技術の基盤となっている [1]。

こうした社会的背景の中で、私たち学生にとって、データに基づいて仮説を立てる力や、データから価値を見出す力、課題に対して仮説を立て、検証し、改善していく力が今後重要となる。特に、機械学習は統計やプログラミングの知識を総合的に応用する実践的な技術であり、単なる理論の理解にとどまらず、現実の問題にどう適用できるかを学ぶことが重要となる。

このような背景を踏まえ、私たちは機械学習のコンペティションに参加し、実践の中でチームで機械学習の技術と運用のプロセスに関するスキルを身につけながら、コンペティションに貢献することを目的として、プロジェクト学習のテーマとして Kaggle への挑戦を選んだ。Kaggle は、実際の企業や研究機関が提供するデータセットを用いた課題に対して、世界中の参加者と競いながら機械学習に関してのスキルを高められ、積極的な参加がコンペティション全体の貢献へもつながるプラットフォームである。

私たちは、この Kaggle という実践の場を通じて、機械学習分野の知識を深めるだけでなく、チームで協働しながら問題解決に取り組む力や、自らのアイデアを具現化する能力も養いたいと考えた。

1.2　機械学習の基礎

機械学習とは、データから規則性を学習し、それをもとに予測や分類を行う技術である。画像認識、音声認識、自然言語処理、ロボット制御など多くの応用分野においての機能を実現するための基幹技術であり、AI の中核的な技術と位置づけられている [1]。

機械学習は大きく分けて、教師あり学習・教師なし学習・強化学習の 3 つに分類される。教師あり学習は、正解が与えられた学習データを使い学習し、未知のデータに対する予測や分類を行う手法である。教師なし学習では、明示的な正解ラベルを与えず、データセット自体が持つ潜在的な構造や統計的パターンを抽出してクラスタリングや次元圧縮を行う手法である。強化学習は、学習主体が環境との相互作用を通じて得られる報酬を最大化するように、試行錯誤的に最適な行動指針を学習する手法である。

機械学習には、線形回帰、決定木、サポートベクターマシン、ニューラルネットワークなど多様なモデルがあり、課題の特性やデータの構造に応じて選択される。また、過学習の防止や汎化性能の向上のために、正則化や交差検証といった技術も重要な役割を果たす。

1.3　Kaggle の基礎

Kaggle は、機械学習を用いた問題解決を実践的に学ぶことができるプラットフォームであり、世界中のユーザーが公開されたデータセットや課題に自由に取り組むことができる。私たちは、ま

ずその基礎を身につけるために、『Python で始める Kaggle スタートブック』[2] を活用しながら読み会形式で学習を進めた。

この書籍を通して、機械学習における主要な手法である分類と回帰の考え方を理解した。分類問題では、探索的データ分析によって仮説を立てた上での特徴量の作成や可視化、特徴量からラベルを予測する手法としてロジスティック回帰や決定木を学び、回帰問題では線形回帰や勾配ブースティングなどを用いて、連続値を予測するアプローチを学んだ。

また、モデルの評価においては、単に学習データへの適合度を見るだけでは不十分であることを知り、交差検証によって汎化性能を評価することの重要性を学んだ。特に、学習データに対して高い精度を示していても、未知のデータには対応できない過学習の問題を回避するためには、検証用データを用いて交差検証を行うことや、正則化、特徴量の選択、モデルの複雑さを抑える工夫が必要であることを実践を通じて理解した。

1.4 講義とのつながり

本プロジェクトでの Kaggle を活用した機械学習を用いた実践的なモデル構築は、大学の講義で学んだ知識と深く関係している。特に「データサイエンス入門」、「データサイエンス基礎」や「機械学習 II」で得た基礎的な内容が、実践の場で役立った。

「データサイエンス入門」と「データサイエンス基礎」では、Python の基本的な文法やデータ分析に用いるライブラリ（Pandas、NumPy、Matplotlib など）の使い方を学んだ。これにより、Kaggle におけるデータの前処理や可視化を行う際に、Python を効率よく扱うことができた。特に、欠損値処理やカテゴリ変数の変換、統計量の計算などは、講義で学んだスキルが基盤となっている。

また、「機械学習 II」では、機械学習モデルの種類やアルゴリズムの特性について学んだ。例えば、決定木系モデルや線形回帰モデルは解釈性に優れていることなど、モデルの利点と欠点を理解することができた。これらの知識は、実際に Kaggle で複数のモデルを試し、比較・評価を行う際の判断材料となった。

このように、講義での理論的な学びを実際のデータに適用することで、知識がより定着し、応用力が身についたと感じている。また、個人が得た知識を共有する場を設け、活発な意見交換を行い、試行錯誤を重ねることで、チームワークの実践を行えた。このように、講義とチームでの実践を行き来することで、単なる知識の習得にとどまらず、チームでの問題解決力の向上にもつながっている。

第 2 章　前期の活動と後期の概要

2.1　入門コンペティション

前期の前半では、実践的な機械学習の基礎を習得することを目的として、複数のチームに分かれ、Kaggle 上で公開されている入門的なコンペティション 4 件に取り組んだ。分類問題、回帰問題、自然言語処理など、異なるタスクを通じて、データ前処理からモデル構築、評価までの一連の流れを実践的に学習した。

Spaceship Titanic

乗客の年齢や客室番号といった属性情報から、異空間に転送されたかを予測する分類問題に取り組んだ。特徴量エンジニアリングや欠損値補完などの前処理を重点的に行い、データ加工がモデル性能に大きく影響することを確認した。最終的な正しく予測されたラベルの割合は 0.79/1.00 であった。

House Prices - Advanced Regression Techniques -

敷地面積や地下室の有無といった住宅情報から、住宅価格を予測する回帰問題に取り組み、多数の特徴量を扱う際の前処理手法や特徴量選択の重要性を学んだ。また、複数モデルを組み合わせるアンサンブル手法を導入し、単一モデルより高い予測性能が得られることを確認した。二乗平均平方根誤差を指標とした最終的な精度は 0.14 であった。

Natural Language Processing with Disaster Tweets

ツイート文が災害に関する内容かを判定する自然言語処理の分類問題に取り組んだ。トークン化や Term Frequency-Inverse Document Frequency を用いたテキストの数値化を行い、テキストデータを機械学習モデルに適用する基礎的手法を習得した。F1 スコアを指標とした、最終的な精度は 0.83/1.00 であった。

Sentiment Analysis on Movie Reviews

映画レビューのフレーズを、否定的から肯定的の 5 段階のの感情ラベルに分類する問題において、Bidirectional Encoder Representations from Transformers[3] および Long Short-Term Memory を用いたモデル構築を行った。両手法の特性を学ぶことで、自然言語処理について理解を待深めた。最終的な正しく予測されたラベルの割合は 0.64/1.00 であった。

2.2　前期本番コンペティション

2.2.1　DRW - Crypto Market Prediction -

コンペティションの概要

本コンペティションでは、暗号資産の市場データを用いて、次の 1 分間における価格変動を予測するモデルの構築を行う。学習データには、取引量などの基本的な特徴量に加え、意味が非公開の

890 個の匿名化特徴量が含まれており、これらの扱いが性能向上の重要な要素となる。

動機と目的

前半に取り組んだ回帰課題で得た、多数の特徴量を整理・選別する経験を、より実践的な場面で応用することを目的として本コンペティションを選択した。評価指標である相関係数の向上を目指し、コンペティションの成果に貢献することを目標とした。

結果と考察

初期段階では、匿名化されていない特徴量のみを用いて LightGBM による学習を行ったが、相関係数は最大でも 0.03 にとどまった。これは、価格変動に強く関与すると考えられる匿名化特徴量を十分に活用できていないことが主な要因であると考えられる。そこで、匿名化特徴量と目的変数との相関分析を行い、相関の高い特徴量を抽出するとともに、特徴量間の相関を可視化することで冗長性の検討を行った。これにより、有効な特徴量を選別する必要性が明確となった。

2.2.2 CMI - Detect Behavior with Sensor Data

コンペティションの概要

本コンペティションでは、手首装着型センサーから取得された時系列データを用いて、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モデルの構築を行う。評価指標には、バイナリ F1 スコアとマクロ F1 スコアを組み合わせた指標が用いられる。

動機と目的

医療・福祉分野における社会的意義の高い課題に対し、機械学習を応用することを目的として本課題に取り組んだ。また、時系列センサーデータの解析を通じて、実践的な機械学習技術の習得を目指した。

結果と考察

角速度および加速度を含む慣性計測ユニットのデータに着目し、これを中心としたベースラインモデルを構築した。XGBoost および 1 次元 CNN を用いて評価を行った結果、最終的に 66% の正解率を得た。特に 1 次元 CNN は、時系列データの局所的なパターンを捉えやすく、XGBoost よりも高い性能を示した。一方で、温度センサーや距離センサーといった他のセンサーデータを十分に活用できておらず、有効な情報を取り逃している可能性がある。また、入力系列の分割方法やモデル構造の設計にも改善の余地が残されている。

2.3 後期の概要

後期においては、2 つのチームに分かれ、前期で得た経験を活かす本番として、コンペティションに取り組んだ。一つ目のチームは Thermophysical Property: Melting Point に、二つ目のチームは ARC(Abstraction and Reasoning Corpus) Prize 2025 に取り組んだ。両チームとも、前期の活動を通じて得られた知識および経験を活用し、学習と試行錯誤を重ねながら課題に取り組んだ。その結果、いずれのチームもコンペティションにおいて上位 10 %以内の予測精度を出すことができた。

第 3 章　Thermophysical Property: Melting Point

3.1　コンペティションの概要・背景

Kaggle 上で公開されている Thermophysical Property: Melting Point というコンペティションを題材とし、有機化合物の分子記述子を用いて融点を予測する機械学習モデルの構築に取り組んだ。

有機化合物の融点を予測することは、化学および化学工学分野における長年の課題の一つである。融点は、薬剤設計、材料選定、化学プロセスの安全性評価などにおいて重要な物性値である一方、その実験的測定には多大な時間やコストを要する。また、物質によっては実験自体が困難、あるいは不可能な場合も存在する [5]。このような背景から、有機化合物の分子記述子をもとに機械学習を用いて融点を予測する手法が近年注目されている。

3.2　目的

本コンペティションの目的は、与えられた学習データをもとに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できる手法を検討することである。

学習データには、説明変数として SMILES（Simplified Molecular Input Line Entry System）や 1 から N まである Group、ID、が与えられ、目的変数として Tm が与えられた。SMILES とは、現代の化学情報処理のために設計された化学表記法体系であり [6]、有機化合物の化学構造式が文字列で表現されている。Group は、Group1 から GroupN まで存在しており、有機化合物に対して付与された特徴量記述子である。ID は、各サンプルを一意に識別する番号である。我々はこれらの説明変数を使用し融点である目的変数の Tm を予測する。

構築したモデルの性能は、コンペティションで定められた評価指標に基づいて評価した。本コンペティションでは平均絶対誤差（Mean Absolute Error：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、回帰モデルの構築方法、モデル評価の考え方を理解するとともに、チームでの分析・検討を通じた実践的なデータ分析能力の向上を目指した。

3.3　手法

3.3.1　RDkit を用いた特徴量エンジニアリング

RDKit の位置づけと本コンペティションにおける役割

本コンペティションでは、有機化合物の融点を機械学習により予測するため、SMILES 文字列で与えられた分子構造情報を数値特徴量へ変換する前処理ツールとして RDKit（Representation of Descriptors Kit）を使用した。RDKit は、分子構造の取り扱い、分子記述子計算、フィンガープリント生成などの機能を提供する、ケモインフォマティクス分野における代表的なオープンソースライブラリである [7]。

本コンペティションにおいて RDKit は予測モデルそのものではなく、分子構造情報を機械学習モデルが扱える数値表現へ変換するための特徴量生成基盤として位置づけられる。

SMILES から分子構造オブジェクトへの変換

データセットに含まれる分子情報は SMILES 文字列として与えられているため、RDKit の Chem.MolFromSmiles 関数を用いて SMILES を分子構造オブジェクトへ変換した。分子構造オブジェクトは、原子の種類、結合関係、結合次数、分岐構造および環構造といった分子構造情報を内部に保持しており、以降の分子記述子計算やフィンガープリント生成はすべてこのオブジェクトを基に行われた。SMILES の解釈に失敗し分子構造オブジェクトが生成できなかった分子については、特徴量を欠損値として扱い、学習データから除外している。

分子記述子による物理化学的特徴量の生成

RDKit による特徴量生成の第一の要素として、本活動では分子記述子を用いた。RDKit が標準で提供する `Descriptors._descList` に基づき、208 種類すべての記述子を算出した。算出された記述子の主なカテゴリーは以下の通りである。

- **物理化学的性質**: MolWt、MolLogP(疎水性)、TPSA（極性表面積）など
- **構造・形状指標**: RingCount、FractionCSP3、Kappa 指数、BCUT2D 系など
- **電子状態**: NumValenceElectrons、PEOE_VSA 系、EState_VSA 系など
- **官能基カウント**: fr_系記述子（アルデヒド、アミド、カルボン酸、ニトロ基、芳香環、硫黄含有官能基等の出現回数）

本活動では、追加の SMILES パターンによる特徴量導入は行わず、RDKit 標準のフラグメント記述子のみを使用した。

MACCS Keys による官能基レベルの構造表現

分子の官能基構成を明示的に表現するため、MACCS(Molecular ACCess System) Keys フィンガープリントを導入した。MACCS Keys は 167 次元の二値特徴量から構成され、事前に定義された官能基や構造モチーフの有無を表す。本活動では RDKit の GenMACCSKeys を用いて各分子の MACCS Keys フィンガープリントを生成し、分子記述子と併せて学習用特徴量に含めた。

Morgan フィンガープリントによる局所構造の表現

さらに詳細な分子構造情報を捉えるため、Morgan フィンガープリント（Extended-Connectivity Fingerprints）を使用した。Morgan フィンガープリントは、各原子を中心とした局所的な部分構造を、指定した結合距離（半径）まで展開し、それらをハッシュ化して二値ベクトルとして表現する手法である。本活動では半径 2、3、4 の三種類の Morgan フィンガープリントを採用し、それぞれ 512 ビットで表現した。これにより、官能基近傍の局所構造から、より広い分子構造文脈までを多段階で特徴量として捉えることが可能となった。

特徴量全体の構成

以上の処理により、本活動では各分子から合計 1911 次元の特徴量を生成した。内訳としては、RDKit が標準で提供する分子記述子 208 次元に加え、官能基や部分構造の有無を表現する MACCS Keys フィンガープリント 167 次元、ならびに分子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 の三条件で算出し、それぞれ 512 次元とした。これらを結合することで、分子の物理化学的性質、官能基構成、局所的および中距離の構造情報を包括的に表現する高次元特徴量空間を構築した。

3.3.2 Lasso 回帰を用いた特徴量選定

学習データの前処理と特徴量選択

RDKit による特徴量生成の結果、各分子から合計 1,911 次元の特徴量を構築した。内訳は、分子記述子 208 次元、MACCS Keys フィンガープリント 167 次元、Morgan フィンガープリント（半径 2、3、4 の各 512 次元）であり、分子の物理化学的性質、官能基構成、ならびに局所構造情報を多角的に数値化した高次元特徴量である。一方で、学習データ数に対して特徴量数が多く、冗長な特徴量や融点予測に寄与しない特徴量が含まれる可能性がある。

そこで本活動では、予測性能の向上およびモデルの汎化性能確保を目的として、複数の特徴量選択手法の比較検討を行った。具体的には、RFE（Recursive Feature Elimination）、KBest（SelectKBest）、および LASSO 回帰（Least Absolute Shrinkage and Selection Operator）を用いて特徴量選択を実施し、交差検証による予測精度の比較を行った。表 3.1 が各手法で選択された特徴量を用いて、5 つの決定木モデルで学習した際の MAE を比較した表である。

本活動で扱う特徴量およびモデル構成に対しては、線形 LASSO 回帰を用いた場合に最も高い予測精度が得られたため、以降の解析では LASSO 回帰による特徴量選択を採用した。LASSO 回帰は、回帰係数に L1 正則化項を加えることで一部の係数を厳密に 0 とする性質を持ち、高次元データにおいて変数選択とモデルの簡素化を同時に行うことが可能である。

特徴量選択に先立ち、連続値特徴量と二値特徴量が混在していることによるスケール差の影響を避けるため、すべての特徴量に対して標準化処理を行った。その後、5 分割交差検証により正則化強度を自動的に決定する Lasso Cross-Validation を用いて学習を行い、回帰係数が 0 でない特徴量のみを選択した。

その結果、1, 911 次元の特徴量のうち 347 次元が選択された。LASSO 回帰の係数の絶対値にに基づく特徴量上位 15 個は図 3.2 のとおりである。全体の内訳としては、分子記述子が 60 次元、MACCS Keys が 46 次元、Morgan フィンガープリントが半径 2、3、4 でそれぞれ 107、66、68 次元であった。分子記述子には、分子サイズや質量を表す HeavyAtomMolWt、電子状態や電荷分布を反映する EState 系および VSA_EState 系記述子、分子トポロジーを表す BalabanJ、Chi3n、Ipc、ならびに疎水性や極性分布を反映する PEOE_VSA 系、SMR_VSA 系、SlogP_VSA 系記述子が含まれていた。

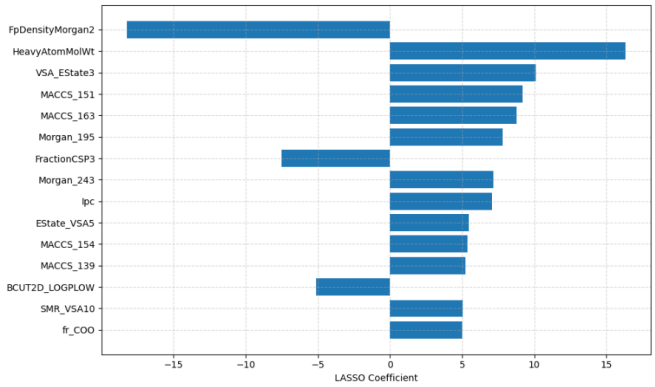

図 3.1　LASSO 回帰によって選択された特徴量上位 15 項目（係数の絶対値順）

また、官能基や部分構造の存在を表すフラグメント系記述子（fr_ 系）も複数選択されており、カルボン酸、アミド、含窒素複素環、硫黄含有官能基など、分子間相互作用や結晶構造に影響を与えると考えられる構造要素が保持されていた。これらの結果は、融点予測において分子の大きさや剛直性、極性分布、官能基構成といった複数の要因が同時に重要である可能性を示唆している。

以上より、LASSO 回帰による特徴量選択は、本活動において高次元な化学構造特徴量から融点予測に有効な情報を抽出する前処理として有効に機能したと結論づけられる。

3.3.3　選定された特徴量を用いて複数の決定木系のモデルを構築

決定木を基盤とするアンサンブル学習モデルを対象とし、学習方法や木の構築方法の異なる 5 種類のモデルを使用した。

決定木とは

決定木とは、データを木構造で表現し、入力変数を条件分岐によって階層的に分割し、目的変数を予測する機械学習手法である。決定木は内部ノード、枝、葉ノードから構成される。データは順に根ノードから入力され、条件を評価しながら枝をたどり、到達した葉ノードに対応する出力がモデルの予測結果となる。

1. Random forest

Random Forest とは、複数の決定木を組み合わせたアンサンブル学習手法であり、分類問題では各決定木の予測結果に基づく多数決によって、回帰問題では予測値の平均によって、最終的な出力が決定される。

与えられた学習データから、ブートストラップ標本の生成や各分割点における特徴量のランダム選択を行い、それらを基に各決定木を構築する。次に、ランダム性を含んだ条件の下で多数の決定木を独立に成長させる。入力データが与えられると、すべての決定木の予測の結果を統合することで最終的な出力が得られる。

Random Forest の特徴として、ブートストラップ標本の生成や各分割点における特徴量のランダム選択により、安定したアンサンブル予測が可能となる点、ラベルノイズや外れ値の影響を受けにくい点などが特徴として挙げられる。

2. Extra Trees Regressor

Extra Trees Regressor とは、Random Forest のアンサンブル理論を基盤としつつ、決定木の構築の際に、分割点と属性選択のランダム性を強化し、各決定木の予測値を平均することで最終的な出力を決定する学習法である。

与えられた学習データから各ノードにおいてランダムに選択された特徴量に対し、値域内からランダムな分割点を生成し、それに基づいてデータを分割する。次に、選択された分割を用い、停止条件まで再帰的に分割を行う。入力データが与えられると、各決定木がそれぞれ独立に数値予測を行い、すべての予測値を平均することで最終的な出力が得られる。

Extra Trees Regressor の特徴として、特徴量選択、分割点がともにランダムに生成されるためアンサンブル全体として分散を小さくすることが可能である点、分割点の最適化探索を行わないため計算効率が高く、大規模データや高次元データに対して有効である点などが挙げられる。

3. XGboost

XGBoost とは、勾配ブースティング決定木を基盤とし、正則化を導入した目的関数の最小化を通じて決定木を構築するアンサンブル学習法である。

与えられた学習データにそれぞれ初期予測値を設定し、予測値と正解値との差を評価する。この誤差の情報をもとに、モデルの改善を反復し、複数の決定木を段階的に積み重ねることで、全体として高い予測性能を持つモデルを構築する。この際、各決定木の複雑さや寄与の大きさに対して正則化を導入する。新しい入力データが与えられると、すべての決定木がそれぞれ予測を行い、それらの結果を合算することで最終的な出力が得られる。

XGboost の特徴として勾配ブースティングに基づく逐次的学習により、バイアスと分散の低減が可能である点、正則化項を目的関数に含めることにより、過学習を抑制することができる点などが挙げられる。

4. Light GBM

Light GBM とは、勾配ブースティング決定木を基盤とし、大規模データや高次元特徴量に対して高速かつメモリ消費を抑えたアンサンブル学習法である。

与えられた学習データにそれぞれ初期予測値を設定し、予測値と正解値との差を評価する。この誤差の情報をもとに、モデルの改善を反復し、複数の決定木を段階的に積み重ねることで、全体として高い予測性能を持つモデルを構築する。この際、木の成長には、損失関数の減少量が最大となる葉ノードを優先的に分割する Leaf-Wise、分割探索には、連続値特徴量をヒストグラムに量子化した上で分割候補を探索するヒストグラムベースの分割探索を導入する。新しい入力データが与えられると、すべての決定木がそれぞれ予測を行い、それらの結果を合算することで最終的な出力が得られる。

Light GBM の特徴として、 ヒストグラムベースの分割探索による分割探索における計算量とメモリ消費が抑制される点、Leaf-wise の導入による集中的な誤差削減効果の高い分割を行える点などが挙げられる。

5. Cat Boost

Cat Boost とは、勾配ブースティング決定木を基盤とし、学習時に生じる予測のずれを抑制する順序付きブースティングを実装した機械学習手法である。

与えられた学習データをランダムな順序に並べ、この順序に基づいて各データ点のカテゴリ特徴量に対する統計量を計算する。次に、予測結果と正解値との差をもとに、誤差を補正する決定木を逐次的に構築する。この際、対称木が用いられ、各木の同一階層においてはすべてのノードで同じ分割条件が適用される。構築された決定木の出力は、既存のモデルに加算され、この操作を繰り返すことで複数の決定木からなるアンサンブルモデルが形成される。新しい入力データが与えられると、すべての決定木がそれぞれ予測を行い、その予測値を合算することで最終的な出力が得られる。

Cat Boost の特徴として、順序付きブースティングによるターゲットリークによる予測の偏りが抑制される点や対称木の導入による分割の自由度の制限によって、学習が安定化する点などが挙げられる。

3.4　最終的な結果と考察

最終的なモデルは、予測精度の向上と未知のデータに対する汎化性能の確保を目的とし、前章で説明した方法で生成および選定した特徴量を用いる、決定木アルゴリズムをベースとした 5 種類のアンサンブル学習モデルとした。これらを個別に最適化した上で、最終的に重み付け投票（Weighted Voting）によって統合する手法を構築した。

3.4.1　個別モデルの選定とハイパーパラメータ最適化

ベースモデルとして、前章で説明した 5 つのアルゴリズムを用いた。

各モデルの性能を最大限に引き出すため、ベイジアン最適化アルゴリズムを用いたハイパーパラメータ自動最適化フレームワークである Optuna を導入した。各アルゴリズムに対し、学習率、木の深さ、サンプリング比率等のパラメータ探索を行い、検証データにおける MAE を最小化する最適なパラメータセットを特定した。

3.4.2　Optuna を用いた重み付けアンサンブルの構築

各モデルの予測特性（得意とするデータ領域）を相補的に活用するため、単純な算術平均ではなく、各モデルの出力に対して重み付けを行う**加重平均アンサンブル（Weighted Ensemble）**を実装した。

構築したモデルの概観を図 3.2 に示す。

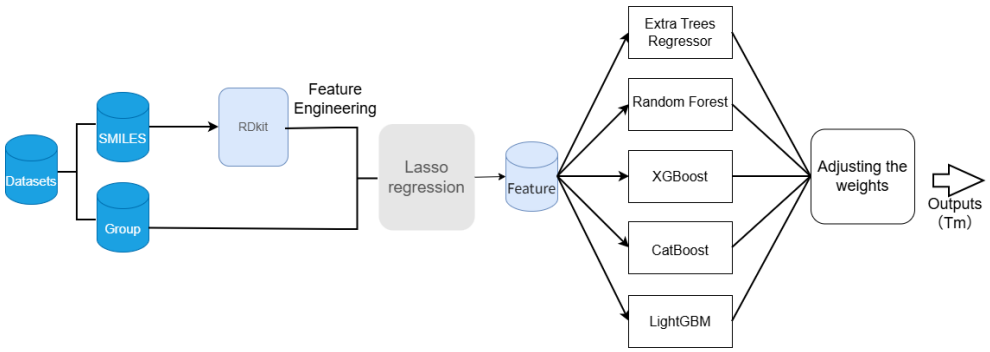

図 3.2　モデルの概観図

重みの決定プロセスにおいては、再度 Optuna を利用した。その結果、単体モデルでは MAE が概ね 26〜28 程度であったのに対し、アンサンブル後には MAE 25.78 を達成した。表 3.1 が単体モデルとアンサンブル後を比較した表である。

3.4.3　結果

上記の最終的なモデルを用いて予測した値をコンペティションに提出した結果、MAE:23.52665 となり、現在 (2025/11/21) のコンペティション内において、上位 10 %以上の精度を持つモデルを構築できた。

表 3.1　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

3.4.4　考察

特徴量とモデルの適合性に関する考察

本モデルが高い精度を達成できた主たる要因は、RDKit を用いて生成した分子記述子とフィンガープリントに対し、LASSO 回帰による適切な特徴量の選択を行った上で、それらの高次元・非線形な関係性を捉えることに長けた決定木系アンサンブルモデルを適用した点にあると考えられる。特に、SMILES データの持つ複雑な情報を、物理化学的な意味を持つ分子記述子やフィンガープリントといった数値特徴量へと翻訳し、それを勾配ブースティングが効率的に学習するというパイプラインが、本課題の性質と合致していたため、最終的に良い結果が得られたと推察される。

深層学習アプローチにおける課題と考察

活動の中盤において、さらなるスコア向上を目指し、1 次元畳み込みニューラルネットワークと Attention 層を組み合わせたモデルの導入を試みた。しかし、結果として MAE は 28 台に留まるなど、決定木系モデルと比較して十分な性能を発揮できなかった。この要因として、今回 CNN に入力したデータが、SMILES 文字列そのものではなく、RDKit によって生成された数値特徴量の羅列であったことが挙げられる。CNN は本来、画像やテキストのような局所的な相関関係を持つデータに対して有効な手法である。しかし、今回生成した特徴量ベクトルにおいては、隣接するカラム同士に構造的な意味や連続性が必ずしも存在しないため、畳み込み層による特徴抽出が有効に機能しなかったと考えられる。CNN を有効活用するためには、SMILES 文字列をトークン化して直接埋め込む手法や、分子をグラフ構造として扱うグラフニューラルネットワークなど、データの構造的特性を維持したまま入力するアプローチが必要であったと推察される。

AutoML による更なる改善の可能性

精度向上の可能性として、AutoGluon などの AutoML ツールの活用が挙げられる。本プロジェクトでは、各モデルの予測値を重み付き平均で統合したが、AutoML ツールが提供するスタッキング技術を用いれば、各モデルの予測値を新たな特徴量とし、別のモデルで学習させる多段構成が可能となる。これにより、各ベースモデルがどのようなケースで予測を誤るかという残差の傾向までをメタモデルが学習できるため、我々が手動で設定した線形結合のアンサンブルよりも、さらに高度な非線形性を捉え、スコアを向上させられる可能性があると考えられる。

第 4 章　ARC Prize 2025

4.1　背景・目的

4.1.1　汎用人工知能（AGI）

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 [6]。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本質である。

4.1.2　現在の AI の限界と ARC テストの意義

近年の大規模言語モデル (Large Language Model: LLM) をはじめとする AI 技術は目覚ましい発展を遂げているものの、その実態は「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 [7]」に依存している側面が強い。そのため、学習データにない「未知の法則」を極めて少ないサンプルから見つけ出す能力、すなわち「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低スコアに留まっている [8]。

4.1.3　ARC Prize 2025 参加の目的

そこで本プロジェクトでは、世界的なコンペティションである「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に挑戦した。具体的には、ARC テストに対応しうる少数のデータから普遍的なルールを抽出するためのアルゴリズム検討および実装を試みた。。

4.2　コンペティションの概要

本コンペティションでは、抽象的な推論能力を測定する「ARC テスト」の正解率を競う。データセット内の各タスクは、数組の入力画像（Input）と出力画像（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、テスト用の入力画像に対する正確な出力画像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。図 4.1 にタスクを視覚化した図を示す。表 4.2 が与えられるデータである。また、本コンペティションには作成したプログラムの実行を 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。

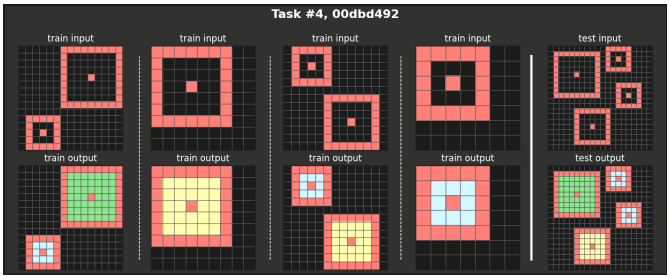

図 4.1　タスクをビジュアル化した図 (Allegich [8] を用い作成)

表 4.1　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	非公開

4.3　手法

4.3.1　特化型ソルバーの構築

本手法では、ARC-AGI-2 の 学習用タスク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の分類を試みた。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。これにより、単純な解法の言語化だけではタスクを分類することは困難であるという課題が明らかになった。そのため、すべてのタスクを分類することは断念し、特徴が明確に抽出できたカテゴリに焦点を絞る方針をとった。具体的には、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。

4.3.2　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[9] 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。具体的には、入力文字列から正解文字列を正確に予測できるようモデルを最適化した。図 4.2 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。

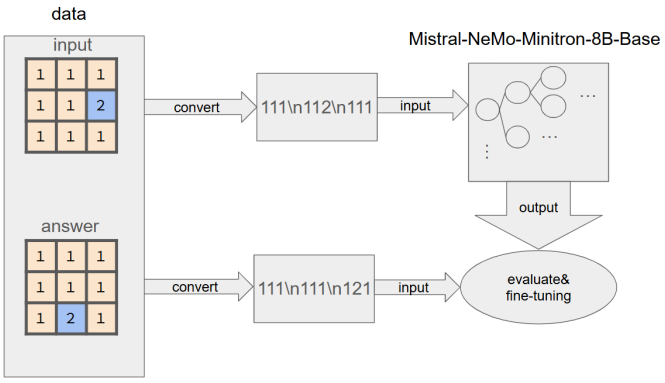

図 4.2　モデルの全体図

4.4　結果と考察

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。

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。この結果は、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を示唆する。

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。これに加えて、事前学習されたデータセット内の統計的なパターンも有効であったと考えらえる。また、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。

第 5 章　総括

5.1　プロジェクトの成果

本プロジェクトでは、1 年を通じて機械学習の理論習得から実践的なコンペティションへの挑戦までを行った。前期においては、入門的なコンペティションを通じて、データの前処理からモデル構築、評価に至る一連のワークフローを習得した。

後期には、専門性の高い課題に対し以下の成果を得ることができた。

- **Thermophysical Property Melting Point**: RDKit を用いた 1911 次元の特徴量生成と Lasso 回帰による選定を実施した。5 種類の決定木系モデルを用いたアンサンブル学習により、MAE 23.52665 という高精度を記録し、上位 10% 以上の成績を収めた。
- **ARC Prize 2025**: AGI の推論能力を測定する ARC テストに対し、KMeans 法によるクラスタリングや LLM を用いた推論を試みた。結果、本グループもコンペティションにおいて上位 10% 以内の成果を達成した。

5.2　講義内容の実践と深化

本プロジェクトは、大学の講義で得た理論を現実のデータに適用する貴重な機会となった。

- 「データサイエンス入門」および「基礎」で学んだ Python ライブラリ（Pandas, NumPy 等）は、欠損値処理や統計計算などのデータ前処理の基盤となった。
- 「機械学習 II」で学んだアルゴリズムの特性や、過学習を防ぐための交差検証・正則化の知識は、モデル選定や評価の判断材料として活用された。

5.3　結論

本実習を通じて、データに基づき仮説を立てて改善する力や、チームで協働する能力を養うことができた。特に、後期に両チームが上位 10% 以内のスコアを達成したことは、継続的な試行錯誤の成果と言える。

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] 我妻幸長：『BERT 実践入門 ― PyTorch + Google Colaboratory で学ぶあたらしい自然言語処理技術』，翔永社，2023.

[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：“Thermophysical Property: Melting Point”，https://www.kaggle.com/competitions/melting-point/overview，（最終アクセス日：2026 年 1 月 9 日）.

[6] David Weininger：SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules，Journal of Chemical Information and Computer Sciences，Vol.28，No.1，pp.31–36，1988.

[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] allegich：“arc-agi-2025-visualization-all-1000-120-tasks”,Kaggle,https://www.kaggle.com/code/allegich/arc-agi-2025-visualizationall-1000-120-tasks,（参照日：2026 年 1 月 14 日）.