# データサイエンスを通して学ぶ課題解決に有効な能力の育成 〜決定木 AI を用いた仮説生成の試み〜

# 後藤貴裕

## 東京学芸大学附属国際中等教育学校

gotoh@u-gakugei.ac.jp

データサイエンスに関する学習を通して得られた、分類(決定木)AIアルゴリズムの手法およびその考え方は、大量の統計データから特定の問題(課題)に強く影響を及ぼしている要因を見出すことを容易にし、さまざまな統計データから問題発見(仮説生成)するための有効なツールとなりうることが期待される。データサイエンスに関する学びが、さまざまな課題解決に有効であることに気づかせる教材としてその具体を提案し、教材研究に関する議論を深めたい。

## 1. はじめに

AI が普通に使える今日の社会では、さまざまなデータから価値を生み出す方法としてデータサイエンスの需要が高まっている.日本学術会議数理科学委員会数学教育分科会(2020)は、氾濫する情報から、批判的にデータの質を評価したり、機械が行う判断の論理の基礎を数学的に理解したりしながら、適切に意思決定や行動に繋げられるようにすることが必要であると提言している.

また、初中等教育の現場では、総合的な探究の時間の探究活動や理数探究などの課題研究などで、課題(問題)解決のための生徒の探究活動が引き続きもとめられている.

本教材は、教科「情報」に関わるトレンドとしてデータサイエンスおよび AI をとりあげ、その手法や考え方を、さまざまな分野の課題解決で活用できる汎化スキルとして位置付け、AI を活用したプログラミングの実習を通して、問題解決のスタートとなる問題発見(課題研究にとっては大事な過程であるが、現場では苦心している)や仮説生成につながる展開となっている.

本教材は、企画・計画の段階であり、授業実践には至っていないが、その具体を提案し、教科「情報」の教材研究の討論の題材としたい.

#### 2. 教材設計とその展開

# 2.1 教材の概要

データサイエンスにおける分類 AI を用いた 機械学習アリゴリズムの学習(実習)を通して 得た技能や考え方を転移させることで,統計データから分類(決定木) AI を用いて特徴量重要 度を明らかにし,対象の状態(ここでは2値) を決定(予測)するために強く関与している説 明変数を特定し,それを課題解決のための仮説 生成に役立てようとするものである.

## 2.2 教育課程における位置付け

教科「情報」の学校設定科目「インフォマティクス」(高校2学年・2単位科目として開設)の教材として開発. 学習指導要領においては,次の複数の科目の内容と深い関わりがある.

#### ○情報「情報Ⅱ」:

(3) 情報データサイエンス (多様かつ大量の データを活用することの有用性に着目し、デ ータサイエンスの手法によりデータを分析し、 その結果を読み取り解釈する活動を行う)

### ○理数「理数探究基礎」:

様々な事象に関わり、理数的な見方・考え 方を用いて、探究の過程を通して、課題を解 決するために必要な資質・能力を身につける。 (オ)事象を分析するための技能〈知識・技能〉 (イ)数学的な手法や科学的な手法などを用い て、探究過程を遂行する力〈思考力・判断力・ 表現力〉

# 2.3 教材のねらい

- 開いた探究や生徒の課題研究において、探究課題や問いの設定のためのツールとして、 分類 AI(決定木)を用いて仮説生成・課題発見をおこなう。
- (オープン) データや統計データから課題を発見する. データを基にして結論(決定)と関係の強い説明変数を特定して, そこから仮説を生成する. その後, 別途課題研究などにおいて仮説に基づく研究を展開したり, 仮説検証を行なったりする.
- 分類 AI (決定木) を Python でプログラミングする. また分類 AI の過学習を回避し 汎化性能を高めるため 「枝刈り (Pruning)」 を行うことを通し機械学習の特性を実践的 に学ぶ.

## 2.4 教材の展開

- ① 統計データを決定木(2値分類) AI で分析・評価できるようにスクレイピングする.
- ◆ 事例として対象とした統計データ
  - A) 教育用標準データセット (SSDSE) 1741 市区町村×多分野 125 項目(全国の全市区町 村の,人口,経済,教育,労働,医療,福 祉など,様々な分野の統計データを収録)
  - B) 天気予報データ(週間予報(7日間)・観測 項目 36 データ) 2006—2015 本稿では A) 教育用標準データセット (SSDSE) を取り上げ, 教材事例とする.

SSDSE データセットを分析することで、社会課題 との関係性の高い要因(説明変数)を調べ、社会課題の解決のための研究課題の仮説生成を行う.

#### テーマ設定

テーマ:独居老人世帯が多くなる条件とは 〔想定する仮説〕65 歳以上世帯員の単独世帯 数は○○の条件のとき多くなる.

分類 AI (決定木) に実装 (分析) し評価できるデータ構造を考えさせるため, 必要に応じて②の実装と往還させる.

- (1) 説明変数を必要に応じて指標化する(絶対数では市町村の規模の影響が大きくなるため、人口あたり、世帯数あたり、等とする).
  - ・#DIV/0!を排除する.
  - ・評価関数の導出に直接関与する変数などは削除する(あきらかに強い関係).
- 明らかに関与しないもの、および0は削除する。
- (2) 評価変数を2値化する.
- 「65 歳以上世帯員の単独世帯数」→2 値化(1,0)で評価する。
- ・総世帯数あたりの「65 歳以上世帯員の単独世帯数」 が 20%を超える→1
- ・総世帯数あたりの「65 歳以上世帯員の単独世帯数」 が 20%を超えない→0

# ② 分類 AI (教師あり学習・決定木) を作成し 実装させる.

 $1741 \text{ rows} \times 123 \text{ columns}$  データのうち 70%を学習データ,残りの 30%をテストデータとするように分割する.

評価変数以外の全ての変数(123)を説明 変数とする多変量解析とする.

2値分類とするため「標準化」はおこなわない. sklearn.tree の機械学習ライブラリを使用して決定木として可視化する.



③ 特徴量重要度を調べ,説明変数を評価し, 仮説生成および課題設定をおこなう.

評価変数の予測に、どの説明変数が AI 予測に強く関与したのかを表す「特徴量重要度」を可視化する. ④と往還しながら⑤に展開する.

- ④ 教師あり学習 AI の評価(学習データとテストデータに対する分類精度の評価)を行い適切な汎化性能について理解する.
  - (1) 学習データとテストデータの分類精度の評価 ( 枝刈りを行っていない状態での) 学習データと テストデータに対する決定木の分類精度を確認 する. 正解率++(test):0.933/(train):0.999
  - (2) 枝刈りをおこなう.学習データとテストデータの分類精度の評価 4層までの深さの決定木 正解率++(test):0.945/(train):0.964

# ⑤ 決定木の意味を読み取り課題解決のための 仮説を生成する

#### 第1層 A130302 <= 0.253 NO

「65歳以上人口(女)の総人口の25.3%以下であれば総世帯数あたりの「65歳以上世帯員の単独世帯数」が20%を超えない」

第2層 A710201<=2.159 NO

「1世帯あたりの一般世帯人員数が 2.195 以下であれば総世帯数あたりの「65 歳以上世帯員の単独世帯数」が 20%を超えない」

第2層 A810105<=0.299 YES

「総世帯数に対して単独世帯数が 29.9 以下であれば総世帯数あたりの「65歳以上世帯員の単独世帯数」が 20%を超える」

〈ゴール〉探究・課題研究の課題(仮説)の 生成(発見)〉課題研究などにおいて仮説検証

## 3. 論点の整理(討論したい課題)

- ・意思決定の分析手法である決定木を逆説的に 決定要因の特定に使用する妥当性
- 特徴量重要度の解釈の妥当性
- ・活用可能な統計データのサンプルの可能性

## 参考文献等

- (1) 日本学術会議数理科学委員会数学教育分科会 (2020):新学習指導要領下での算数・数学教育の円滑の実施に向けた緊急提言:統計教育の 実効性に向けて、日本数学教育学会誌 102 巻 10 号 p. 15-31 (2020)
- (2) 細田幸希:高等学校段階におけるデータサイエンス教育に関する世界的動向-ドイツのProDaBi プロジェクトに関する研究の概観を通じて-、日本科学教育学会第45回年会論文集p.109-112(2021)
- (3) 吉田雅裕: Python で学ぶ初めてのデータサイエンス, 技術評論社 (2023)
- (4) SSDSE (教育用標準データセット): 独立行政法人統計センター(最終確認日:2025.05.30) https://www.nstac.go.jp/use/literacy/ssdse/