探究と情報をつなぐ（４）

～SSH 指定校としての「データ駆動型探究」に向けた実践～

伊藤　大貴
大分県立大分舞鶴高等学校

2024 年 8 月 29 日

１．統計と AI

前回取り上げた「数理・データサイエンス・AI」の「AI」の理解において、統計は重要な位置を占めており、高校「情報Ⅰ」及び「情報Ⅱ」における統計・データサイエンスに関する学習の重要性が示されている。中でも機械学習は、統計の考え方を基盤にしてデータのパターンや法則性を見つけ出し、それをもとに予測や意思決定を行う技術である。高校「情報Ⅰ」および「情報Ⅱ」で学ぶ統計・データサイエンスの内容は、機械学習の基礎的な概念や手法を理解するための重要なステップとして位置づけられている。

たとえば、回帰分析や相関係数の理解は、データ間の関係性を探るための基本的なスキルである。また、データの可視化や確率分布の概念は、機械学習モデルの結果を正しく解釈し、意味ある洞察を引き出すために欠かせない要素であり、これらの知識を活用することで、生徒は単なるデータ分析にとどまらず、モデルの構築や評価に関する初歩的な理解を深めることができる。

さらに、機械学習の実践においては、大量のデータから自動的に学習し、汎化能力を持つモデルを構築する手法が重要になるが、その背後にあるアルゴリズムや統計的な理解があることで、ブラックボックス的に結果を信じるのではなく、理論に基づいた批判的な思考を養うことが可能となる。

例えば、機械学習を応用した技術に、画像認識というものがある。自動運転における画像認識では、カメラで撮影した画像を解析して、歩行者や他の車両を認識することができる。これには、事前に大量のデータ（写真）から学習したモデルが使用されている。つまり、事前に画像から学ぶことで「人」や「車」を識別する力を養い、本番（運用時）では、学んだ力（モデル）を使って「人」や「車」を識別することができる。この技術により、車両は周囲の状況をリアルタイムで把握し、安全な運転を可能にしている。機械学習はこのように、自動運転技術の重要な基盤となっている。また、機械学習は、音声認識や手書き文字認識，ネットショッピングの商品リコメンド，天気予報など、様々な分野においても活用されており、これらの技術を「教養」として学ぶ必要がある。

このように、高校での統計・データサイエンスの学習は、単に数値を扱う技術を身につけるだけでなく、AI が社会に及ぼす影響や応用分野についても深く考えるための土台を築くことにつながる。

しかし、統計と AI の関係性について、実践的に学ぶ教材は存在するが、どの教材において

も、敷居が高い。また、データの前処理から細かなステップをプログラミングや学習アルゴリズムから学ぶには時間がかかるだけでなく、「情報Ⅱ」を開設している学校が少ないことから、多くの生徒が学ぶことは現段階では困難である。

そこで、統計と AI の関係性について学ぶために、身近なデータで簡単に機械学習を行うことのできる Web アプリケーションの開発を行った。

具体的には、Python で Web アプリケーションを開発することのできる Streamlit というフレームワークを使用し、AutoML（自動化された機械学習）ライブラリの PyCaret を用いて機械学習を行うステップを実践的に理解できる工夫を行った。これにより、生徒自身が集めたデータ（オープンデータも可能）を使用し、回帰問題を体験することができる。

機械学習における「回帰問題」とは、連続的な数値データを予測するための問題を指す。これは、入力データ（特徴量）と連続的な出力データ（ターゲット変数）との間の関係をモデル化し、新しいデータに対して数値的な予測を行うものである。

例えば、以下のようなデータを利用することができる。

特徴量（入力）:　　住宅の面積・部屋の数・築年数・最寄り駅からの距離

ターゲット変数（出力）:　　住宅の販売価格

機械学習では、特徴量からターゲット変数を予測するモデルを作成することができる。つまり、過去のデータから特徴量と価格の関係を学習し、新しい物件の価格を推定することができるのである。また、学習を行う段階で、「どの変数が影響を及ぼしているのか」などを確認することができるため、分析として活用することもできる。

機械学習を行う目的は、大量のデータからパターンや関係性を自動的に学習や分析を行いし、将来のデータや未知のデータに対して正確な予測や意思決定を行うことである。人間が手作業で分析するには限界があり、複雑なデータや大量のデータを効率的に処理するために機械学習が活用される。

２．機械学習 Web アプリケーション「easyAutoML」

主な条件は以下の通りである。

①タブレット PC で軽快に動作することができる

②データの加工に関する操作を最小限に抑える

③タッチ操作のみで操作することができる

④機械学習のステップを理解できる

①②③に関しては、「easyStat」と同様、時間的・空間的・技能的な制約に依存しない配慮を行った。そのため、機械学習＋Web アプリケーションという観点から、使用言語は Python を選定し、Streamlit 及び PyCaret ライブラリを用いて、開発を行った。④については、一般的な機械学習の流れ（前処理、モデル比較・学習、チューニング・評価）の理解を促すために、ステップごとにボタンを実装し、使用者の理解を補助する工夫を行った。また、モデルのダウンロードや、モデルの可視化機能の実装も行った。特に、「特徴量の重要度」については、ターゲット変数への影響量を変数ごとに表示しているため、高校生の探究における分析

手法の一つとして使用することが考えられる。

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
et	Extra Trees Regressor	0.0759	0.0847	0.2423	0.97	0.0367	0.014	0.159
gbr	Gradient Boosting Regressor	0.0852	0.0836	0.2453	0.9671	0.0398	0.0163	0.124
rf	Random Forest Regressor	0.1112	0.1038	0.2971	0.9609	0.0477	0.0212	0.191
xgboost	Extreme Gradient Boosting	0.0863	0.1258	0.2834	0.9553	0.0426	0.0153	0.134
dt	Decision Tree Regressor	0.0729	0.1345	0.2846	0.9544	0.0433	0.0131	0.101
lightgbm	Light Gradient Boosting Machine	0.1426	0.1245	0.3269	0.952	0.0558	0.0284	0.116
ada	AdaBoost Regressor	0.3021	0.1775	0.4094	0.9358	0.0747	0.0688	0.132
ridge	Ridge Regression	0.2978	0.3889	0.5687	0.8741	0.0862	0.0614	0.103
br	Bayesian Ridge	0.3006	0.398	0.5736	0.8712	0.0868	0.0619	0.1
lr	Linear Regression	0.3082	0.4262	0.589	0.8619	0.0885	0.0631	0.824

図 2　「easyAutoML」の操作画面（モデルの比較）

Fold	MAE	MSE	RMSE	R2	RMSLE
0	0.2027	0.0679	0.2605	0.9805	0.0421
1	0.2548	0.1207	0.3475	0.9678	0.0524
2	0.3185	0.2213	0.4704	0.8894	0.0756
3	0.297	0.2313	0.481	0.9253	0.0814
4	0.2664	0.197	0.4439	0.9182	0.0848
5	0.3155	0.228	0.4775	0.93	0.0729
6	0.3112	0.1976	0.4445	0.9252	0.075
7	0.2639	0.1684	0.4104	0.949	0.0781
8	0.223	0.0765	0.2766	0.9766	0.0409
9	0.2742	0.2029	0.4504	0.9212	0.0625

Fold	MAE	MSE	RMSE	R2	RMSLE
0	0.0825	0.0291	0.1705	0.9916	0.0312
1	0.2605	0.2555	0.5055	0.9318	0.0814
2	0.252	0.2212	0.4703	0.8895	0.0745
3	0.2172	0.2806	0.5297	0.9093	0.1002
4	0.1835	0.1585	0.3981	0.9342	0.0707
5	0.2298	0.2151	0.4637	0.934	0.0706
6	0.1893	0.1393	0.3733	0.9472	0.0602
7	0.2457	0.1827	0.4275	0.9447	0.0809
8	0.1229	0.0319	0.1787	0.9902	0.0274
9	0.1747	0.0821	0.2864	0.9681	0.0445

図 3　「easyAutoML」の操作画面（チューニング）

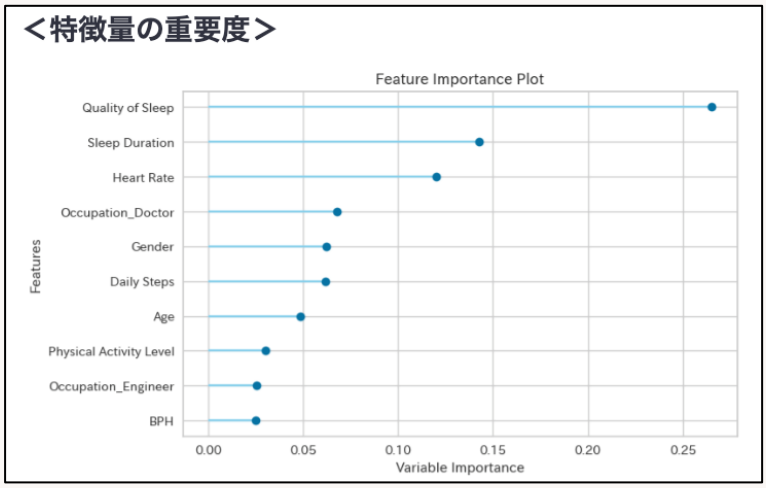

図 4　「easyAutoML」の操作画面（可視化及び評価）

参考：https://huggingface.co/spaces/itou-daiki/pycaret_datascience_streamlit_demo

３．探究をより深めるための成果発表

前回取り上げた「easyStat」や今回紹介する「easyAutoML」等を使って、AI の基礎である統計や機械学習を学びながら探究を行うことができる。実際に、本校では、これらの Web アプリケーションを使って探究を行っている。また、探究の成果として、様々なイベントや学会を通した成果の発表を行うようになった。イベントや学会に参加することで、専門家から助

言を得るだけでなく、他の探究に触れることで、さらなる発展が期待できる。

図５　データサイエンスを活用した課題研究ポスターの例

この探究は、高校生の精神的疲労を可視化し、適切なコーピング方法を提案することを目的に実施したものである。ストレスに関するアンケート調査で得られたデータから、因子分析を行うことにより、「人間関係」「心理的余裕」「食事・睡眠」の３因子を抽出し、「高校生の

ストレス」を定義した。次に、心拍などのバイタルデータを用いて、定義した「高校生のストレス」への影響を機械学習を用いて分析した結果、睡眠や運動が「高校生のストレス」に影響していることが示唆された。さらに、測定結果と対処法を提示する Web アプリケーションを開発し、高校生が自身のストレスを管理・軽減できるよう支援する実用的なツールを提供した。

４．科学部情報班の歩み

このような情報技術を用いた探究や、情報学そのものの探究を行う生徒が増えたため、今年度より、「科学部情報班」を設置している。

科学部情報班は自らを「Sc!TechS(Science Club Information Technology Squad)」と名乗り、様々な研究や情報オリンピック（競技プログラミング）に取り組むことで、探究のロールモデルを波及する役目を担っている。

＜主な成果＞

・R6　情報オリンピック（JOI）2024　１次予選　**10 名突破**
・R6　第７回　中高生情報学研究コンテスト　応募中
・R6　第５回　学力向上アプリコンテスト　最優秀賞
・R6　U22 プログラミング・コンテスト 2024　事前審査突破
・R6　AtCoder Junior League　学校対抗 45 位（11/15 時点）
・R5　第６回　中高生情報学研究コンテスト全国大会　奨励賞
・R5　第６回　中高生情報学研究コンテスト予選　優秀賞

部員は探究（研究）を行う中で、自ら問題を発見し、自ら新しい知識を求めながら学びを切り拓いていく様子がうかがえる。世の中を少しでも改善しようとする主体性を発揮する様子を見て、「探究」と「情報」の関わりを再認識することができた。

５．おわりに

１回目の連載で述べたように、「総合的な探究の時間」には「主体性」が重要であり、教員はあくまでもファシリテーションが役目である。生徒が自走するまでの支援が必要であるが、科学的な根拠を抽出するための統計及び機械学習を用いた分析には、２回目、３回目の連載で紹介した Web アプリケーションによって、探究を深めることができると考えられる。

様々な教材や Web アプリケーションを活用した「学びの変革」によって、社会を変えようとする生徒が増えることを願っている。

情報は、探究だけでなく、各教科での学びをつなげる教科であり、今後の子どもたちにとって重要な位置づけである。「共通テスト」にとらわれることなく、この科目の素晴らしさを伝えていくことが、情報科を担当する教員の役目ではないだろうか。

―情に報いる―　大分県立大分舞鶴高等学校　伊藤大貴

引用参考文献

1. Dit-lab. (2024)：easyStat(DEMO)，HuggingFace，
https://huggingface.co/spaces/itou-daiki/easy_stat_demo
(参照日 2024.6.13)
2. Dit-lab. (2024)：easyAutoML(DEMO)，HuggingFace，
https://huggingface.co/spaces/itou-daiki/pycaret_datascience_streamlit_demo
(参照日 2024.6.13)