Connecting Inquiry and Information (4)

~Series: Practices Toward "Data-Driven Inquiry" as an SSH-Designated School~

Daiki Ito
Oita Maizuru High School
29 August 2024

1. Statistics and AI

In the previous discussion on "Mathematics, Data Science, and AI, " statistics play a crucial role in understanding the "AI" component. The importance of learning statistics and data science in high school courses such as "Information I" and "Information II" has been emphasized. Among these, machine learning is a technology that identifies patterns and regularities in data based on statistical principles, enabling predictions and decision-making. The statistical and data science content in "Information I" and "Information II" is positioned as a vital step in understanding fundamental concepts and methods of machine learning.

For example, understanding regression analysis and correlation coefficients are basic skills for exploring relationships between data. Similarly, concepts like data visualization and probability distribution are essential for correctly interpreting machine learning model results and deriving meaningful insights. By leveraging this knowledge, students can go beyond simple data analysis to gain an introductory understanding of model construction and evaluation.

In the practical application of machine learning, the method to build models capable to learn automatically from large datasets with high generalization ability is essential. In addition, understanding the underlying algorithms and statistical concepts allows students to engage in critical thinking based on theory, rather than blindly trusting results produced by a "black box."

One practical example of machine learning is image recognition, as seen in autonomous driving. In such systems, cameras capture images that are then analyzed to recognize pedestrians and other vehicles. This relies on models trained with large datasets (photos) to identify "people" and "vehicles." These trained models are then applied during operation to recognize and react to real-world scenarios. This technology enables vehicles to monitor their surroundings in real-time, ensuring

safer driving. Machine learning thus forms a critical foundation for autonomous driving technologies. Beyond that, machine learning is widely utilized in various fields, such as speech recognition, handwriting recognition, e-commerce product recommendations, and weather forecasting. Understanding these technologies as part of general literacy is increasingly important.

Learning statistics and data science in high school not only equips students with technical skills for handling numbers but also lays the groundwork for deeply considering the societal impacts and applications of AI.

However, while there are materials available for studying the relationship between statistics and AI, many of these are advanced and inaccessible. Learning the detailed steps of data preprocessing, programming, and learning algorithms is time-intensive, and the limited availability of Information II courses in schools makes it difficult for many students to gain this knowledge at present.

To address this, a web application was developed to allow students to easily perform machine learning with familiar datasets. Using the Python-based Streamlit framework for web application development and the AutoML library PyCaret, the application facilitates a practical understanding of the steps involved in machine learning. With this tool, students can use datasets they collect themselves (or open datasets) to experience regression problems.

In machine learning, a "regression problem" refers to predicting continuous numerical data. It involves modeling the relationship between input data (features) and continuous output data (target variables) to make predictions for new data. For example, the following dataset can be used:
Features (Inputs): Housing area, number of rooms, age of the house, distance to the nearest station
Target Variable (Output): House sale price

Machine learning creates models that predict the target variable from the features. By learning from past data, these models can estimate the prices of new properties. Additionally, during the learning phase, students can analyze the data to identify which variables have the most influence, providing insights for analysis.

The goal of machine learning is to automatically learn patterns and relationships

from large datasets, enabling accurate predictions and decisions for future or unknown data. Machine learning is particularly useful for efficiently handling complex or large datasets, which are beyond the limits of manual analysis.

2. Machine Learning Web Application "easyAutoML"

The main requirements for developing this application are as follows:

1. It should operate smoothly on tablet PCs.
2. It should minimize the steps required for data preprocessing.
3. It should allow for touch-only operation.
4. It should enable understanding of machine learning steps.

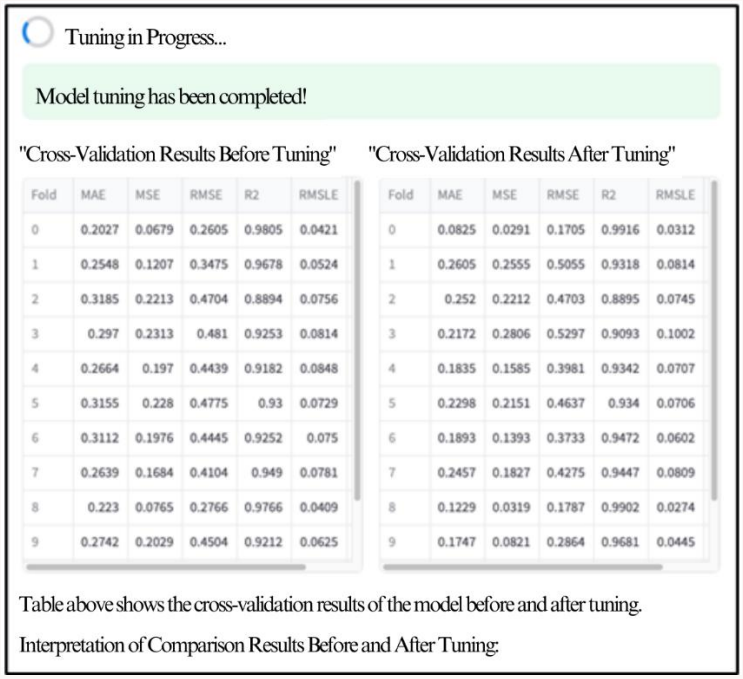
Requirements 1, 2, and 3 were addressed similarly to easyStat, ensuring the application was free from temporal, spatial, and technical constraints. Python was selected as the programming language, and the Streamlit and PyCaret libraries were used for development, emphasizing machine learning and web application perspectives. For requirement 4, buttons were implemented for each step of the machine learning process (preprocessing, model comparison and training, tuning, and evaluation) to aid user understanding. Additional features, such as model download and visualization, were also included. In particular, a feature for "feature importance" was added, displaying the impact of each variable on the target variable. This feature makes the application suitable for use as an analytical method in high school inquiry projects.

Model Comparison Results

The following table shows the performance of each available model.

	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

Figure2: "easyAutoML" Interface (Model Comparison)



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

Figure3: The "easyAutoML" Interface (Tuning)

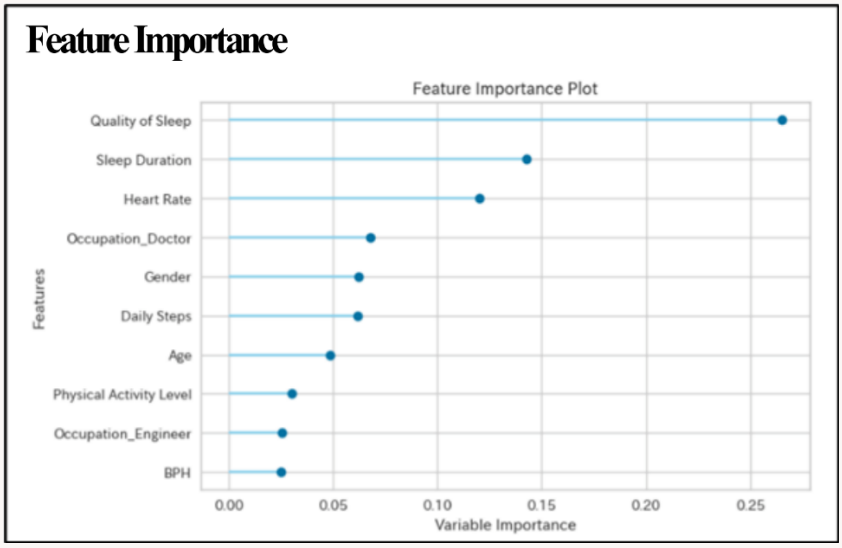


Figure4: The "easyAutoML" Interface (Visualization and Evaluation)

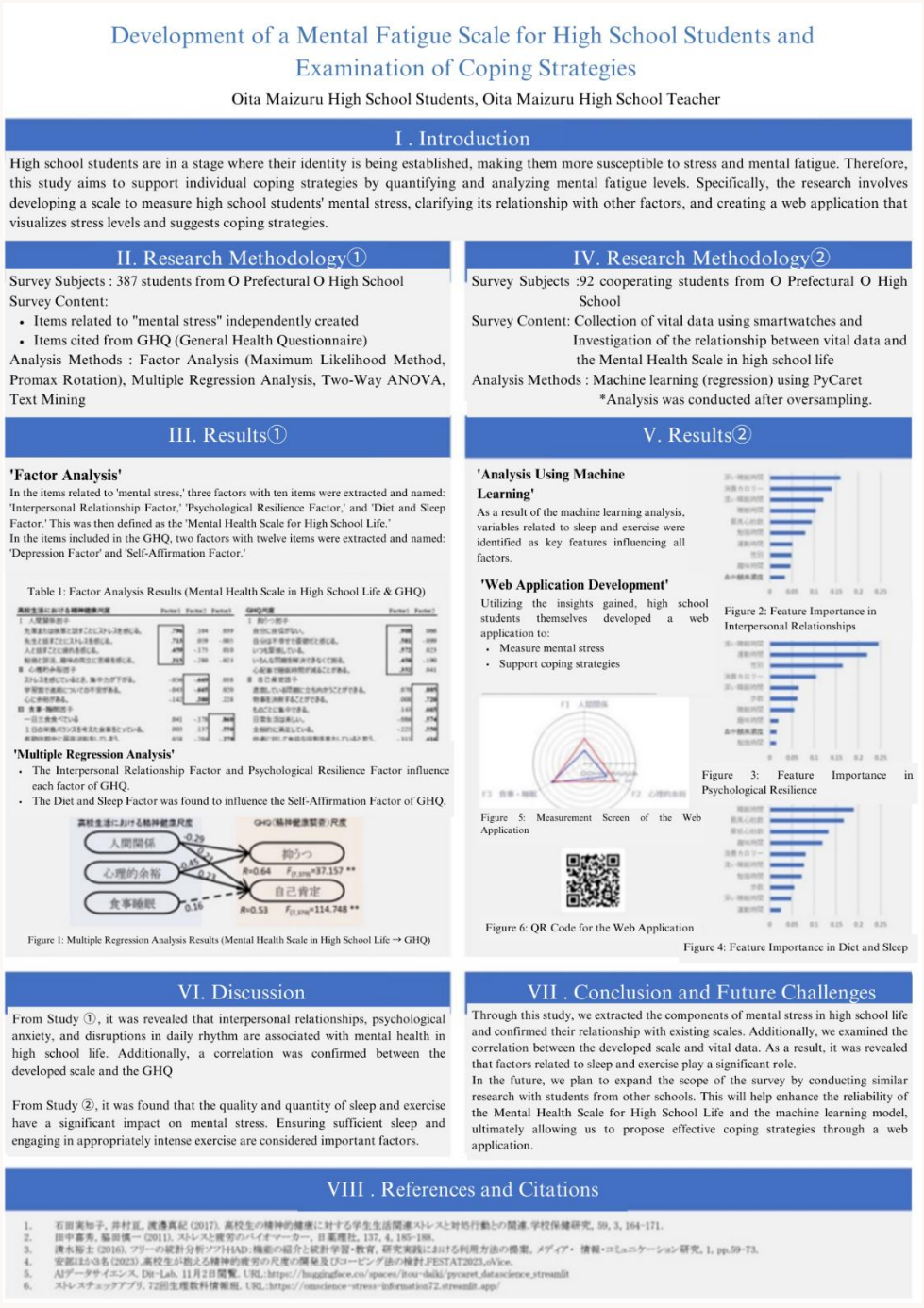
Reference:

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

3. Presenting Results to Deepen Inquiry

Using tools such as easyStat, introduced previously, and easyAutoML, introduced in this section, students can conduct inquiry-based activities while learning the fundamentals of AI, including statistics and machine learning. At our school, these

web applications are actively utilized in inquiry projects. Additionally, as a culmination of their inquiry efforts, students now present their results at various events and academic conferences. Participating in these events and conferences not only allows students to receive advice from experts but also exposes them to other inquiry projects, fostering opportunities for further development and growth.



Development of a Mental Fatigue Scale for High School Students and Examination of Coping Strategies

Oita Maizuru High School Students, Oita Maizuru High School Teacher

I . Introduction

High school students are in a stage where their identity is being established, making them more susceptible to stress and mental fatigue. Therefore, this study aims to support individual coping strategies by quantifying and analyzing mental fatigue levels. Specifically, the research involves developing a scale to measure high school students' mental stress, clarifying its relationship with other factors, and creating a web application that visualizes stress levels and suggests coping strategies.

II. Research Methodology①

Survey Subjects : 387 students from O Prefectural O High School
Survey Content:
- Items related to "mental stress" independently created
- Items cited from GHQ (General Health Questionnaire)

Analysis Methods : Factor Analysis (Maximum Likelihood Method, Promax Rotation), Multiple Regression Analysis, Two-Way ANOVA, Text Mining

III. Results①

'Factor Analysis'

In the items related to 'mental stress,' three factors with ten items were extracted and named: 'Interpersonal Relationship Factor,' 'Psychological Resilience Factor,' and 'Diet and Sleep Factor.' This was then defined as the 'Mental Health Scale for High School Life.'
In the items included in the GHQ, two factors with twelve items were extracted and named: 'Depression Factor' and 'Self-Affirmation Factor.'

Table 1: Factor Analysis Results (Mental Health Scale in High School Life & GHQ)

'Multiple Regression Analysis'

- The Interpersonal Relationship Factor and Psychological Resilience Factor influence each factor of GHQ.
- The Diet and Sleep Factor was found to influence the Self-Affirmation Factor of GHQ.

Figure 1: Multiple Regression Analysis Results (Mental Health Scale in High School Life → GHQ)

IV. Research Methodology②

Survey Subjects :92 cooperating students from O Prefectural O High School
Survey Content: Collection of vital data using smartwatches and Investigation of the relationship between vital data and the Mental Health Scale in high school life
Analysis Methods : Machine learning (regression) using PyCaret
*Analysis was conducted after oversampling.

V. Results②

'Analysis Using Machine Learning'

As a result of the machine learning analysis, variables related to sleep and exercise were identified as key features influencing all factors.

'Web Application Development'

Utilizing the insights gained, high school students themselves developed a web application to:
- Measure mental stress
- Support coping strategies

Figure 2: Feature Importance in Interpersonal Relationships

Figure 3: Feature Importance in Psychological Resilience

Figure 5: Measurement Screen of the Web Application

Figure 6: QR Code for the Web Application

Figure 4: Feature Importance in Diet and Sleep

VI. Discussion

From Study ①, it was revealed that interpersonal relationships, psychological anxiety, and disruptions in daily rhythm are associated with mental health in high school life. Additionally, a correlation was confirmed between the developed scale and the GHQ

From Study ②, it was found that the quality and quantity of sleep and exercise have a significant impact on mental stress. Ensuring sufficient sleep and engaging in appropriately intense exercise are considered important factors.

VII . Conclusion and Future Challenges

Through this study, we extracted the components of mental stress in high school life and confirmed their relationship with existing scales. Additionally, we examined the correlation between the developed scale and vital data. As a result, it was revealed that factors related to sleep and exercise play a significant role.
In the future, we plan to expand the scope of the survey by conducting similar research with students from other schools. This will help enhance the reliability of the Mental Health Scale for High School Life and the machine learning model, ultimately allowing us to propose effective coping strategies through a web application.

VIII . References and Citations

1. 石田実知子, 井村亘, 渡邊真紀 (2017). 高校生の精神的健康に対する学生生活関連ストレスと対処行動との関連. 学校保健研究, 59, 3, 164-171.
2. 田中喜秀, 脇田慎一 (2011). ストレスと疲労のバイオマーカー. 日薬理誌, 137, 4, 185-188.
3. 清水裕士 (2016). フリーの統計分析ソフトHAD: 機能の紹介と統計学習・教育, 研究実践における利用方法の提案. メディア・情報・コミュニケーション研究, 1, pp.59-73.
4. 安部ほか3名 (2023).高校生が抱える精神的疲労の尺度の開発及びコーピング法の検討.FESTAT2023,oVice.
5. AIデータサイエンス. Dit-Lab. 11月2日閲覧. URL:https://huggingface.co/spaces/itou-daiki/pycaret_datascience_streamlit
6. ストレスチェックアプリ, 72回生理教科情報班. URL:https://omscience-stress-information72.streamlit.app/

Figure5: Example of a Research Poster Utilizing Data Science

This inquiry project aimed to visualize the mental fatigue of high school students and propose appropriate coping methods. By conducting factor analysis on data obtained from a stress-related questionnaire survey, three factors were identified: "Interpersonal Relationships", "Psychological Resilience", and "Diet and Sleep"—together defining "high school student stress." Next, using vital data such as heart rate, machine learning was applied to analyze the impact on the defined "high school student stress." The analysis suggested that sleep and physical activity significantly influence stress levels in high school students. Additionally, a web application was developed to present the measurement results and coping strategies, providing a practical tool for students to manage and reduce their stress.

4. Progress of the Science Club Information Team

As more students have begun engaging in inquiry-based projects utilizing information technology, as well as exploring informatics itself, the "Science Club Information Team" was established this academic year. The team, self-named "Sc!TechS (Science Club Information Technology Squad)", undertakes various research projects and participates in events such as the Olympiad in Informatics (competitive programming). They serve as role models, spreading the value of inquiry-based learning.

Major Achievements:

・R6 Olympiad in Informatics (JOI) 2024: 10 students passed the first preliminary round

・R6 7th High School Informatics Research Contest: Currently submitted

・R6 5th Learning Improvement App Contest: Grand Prize

・R6 U22 Programming Contest 2024: Passed the preliminary review

・R6 AtCoder Junior League: 45th place in school competition (as of November 15, 2024)

・R5 6th High School Informatics Research Contest National Tournament: Encouragement Award

・R5 6th High School Informatics Research Contest Preliminary Round: Excellence Award

Through their inquiries and research, team members demonstrate the ability to identify problems independently and actively seek new knowledge, carving their own paths in learning. Their initiative to make even small improvements to society highlights the profound connection between "inquiry" and "information".

5. Conclusion

As discussed in the first article in this series, "autonomy" is crucial in the "Period for Inquiry-Based Cross-Disciplinary Study", and teachers' roles are primarily facilitative. While support is needed to guide students until they become self-reliant, the web applications introduced in the second and third articles can deepen inquiry by enabling statistical and machine learning-based analyses to extract scientific evidence.

It is my hope that the "transformation of learning" through various teaching materials and web applications will inspire more students to aspire to change society.

Information not only supports inquiry but also connects learning across various subjects, making it a pivotal discipline for the next generation. It is the responsibility of teachers in the field of information to convey the value of this subject, going beyond the framework of "standardized tests", and to appeal its importance in shaping the future.

— Repaying with Passion — Hiroki Ito, Oita Maizuru High School, Oita Prefecture

References

1. Dit-lab. (2024): easyStat (DEMO), HuggingFace,
https://huggingface.co/spaces/itou-daiki/easy_stat_demo
(Accessed: June 13, 2024)

2. Dit-lab. (2024): easyAutoML (DEMO), HuggingFace,
https://huggingface.co/spaces/itou-daiki/pycaret_datascience_streamlit_demo
(Accessed: June 13, 2024)