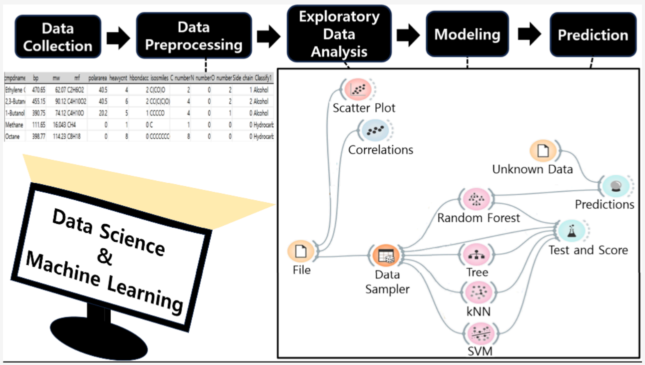
JOURNAL OF
CHEMICAL EDUCATION

Integrating Data Science and Machine Learning to Chemistry Education: Predicting Classification and Boiling Point of Compounds

Published as part of Journal of Chemical Education virtual special issue "Investigating the Uses and Impacts of Generative Artificial Intelligence in Chemistry Education".

Shin-Yu Kim, Inseong Jeon, and Seong-Joo Kang*

Cite This: J. Chem. Educ. 2024, 101, 1771−1776 Read Online



cmpdname	bp	mw	mf	polararea	heavycnt	hbondacc	isosmiles	C number	N number	O number	Side chain	Classify1
Ethylene (	470.65	62.07	C2H6O2	40.5	4	2	C(CO)O	2	0	2	1	Alcohol
2,3-Butan	455.15	90.12	C4H10O2	40.5	6	2	CC(C(C)O)	4	0	2	2	Alcohol
1-Butanol	390.75	74.12	C4H10O	20.2	5	1	CCCCO	4	0	1	0	Alcohol
Methane	111.65	16.043	CH4	0	1	0	C	1	0	0	0	Hydrocarb
Octane	398.77	114.23	C8H18	0	8	0	CCCCCCCC	8	0	0	0	Hydrocarb

ABSTRACT: Artificial intelligence (AI) and data science (DS) are receiving a lot of attention in various fields. In the educational field, the need for education utilizing AI and DS is also being emerged. In this context, we have created an AI/DS integrating program that generates a compound classification/regression model using characteristics of compounds and predicts classification and boiling points of compounds from an unknown dataset. Students have experienced data collection and preprocessing, exploratory data analysis, modeling, and prediction. The No-Code-Low-Code Orange3 tool has been used for the process of modeling and prediction so that even beginners can easily perform Machine Learning (ML) analysis. The raw dataset containing 24 characteristics for 277,569 compounds went through data preprocessing process and became a well-refined dataset. The Random Forest model accurately predicted whether the type of compound in the unknown dataset was hydrocarbons, alcohols, or amines and predicted the boiling points of the some arbitrary compounds within the average error range of 4.49K. This activity will provide meaningful implications for how AI/DS technology could be integrated into each domain.

KEYWORDS: *High School, Introductory Chemistry, Demonstration, Interdisciplinary, Multidisciplinary, Classification, Regression*

■ INTRODUCTION

As digital technology takes root throughout society, humanity is entering an era of digital transformation.[1] The era of digital transformation utilizes digital technologies such as data science (DS) and artificial intelligence (AI), and AI provides insights from big data.[2] In other words, big data is the material for AI, and AI is a tool that converts big data into insights necessary for decision-making.[3]

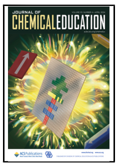
Efforts are being made to add AI to human abilities. MIT president Leo Rafael Reif said "the idea is to use AI, machine learning and data science with other academic disciplines to educate the bilinguals of the future, defining bilingual as those working in biology, chemistry, politics, history and linguistics with computing skills that can be used in their field." He also said "To educate bilinguals, we have to create a new structure."[4,5] In the educational field, there is a need to explore ways to integrate AI, machine learning and data science in each subject.

The integration of AI technology can be classified into three types; to use AI algorithms and models already developed, to develop models using algorithms, and to develop AI algorithms.[6] Recently, AI tools corresponding to the first and second types have appeared. It allows anyone to analyze big data with just drag and drop without coding.[7] Like computer use, AI use has moved from the realm of experts to the realm of the general public. Tools that can be analyzed with just drag and drop are called No-Code-Low-Code (NCLC) tools.[8,9] In this study, Orange3,[10] one of NCLC tools, would be used.

Big data is the material for Machine Learning(ML), and the data preprocessing in the AI and DS processes is an essential and very important step. However, in AI and DS education, students often skip the data preprocessing and proceed with already well-organized dataset due to class time or difficulty.[11] Therefore, this study developed activities that allow students to experience data preprocessing in AI and DS education.

We thought that scientific inquiry activities using AI and DS are necessary. In this study, we presented an ML program that

Received: October 9, 2023
Revised: February 15, 2024
Accepted: February 20, 2024
Published: March 6, 2024



https://doi.org/10.1021/acs.jchemed.3c01040

generates a compound classification/regression model using the chemical and physical properties of hydrocarbon, alcohol, and amine molecules. Additionally, using the created model, students perform inquiry activities to predict the classification and boiling point of arbitrary compounds.

■ HAZARDS

Key concerns include the risk of model overfitting and data bias, which can lead to inaccurate or skewed predictions. Ethical considerations are crucial, focusing on data privacy, informed consent, and the security of sensitive information. The prevention of algorithmic bias is essential to ensure fairness and equality in AI-driven decisions. Transparency in AI models is imperative for accountability.

■ EXPERIMENTAL OVERVIEW

Inquiry activity to classify compounds and predict boiling points of compounds using ML and DS is suitable for high school students or students taking introductory chemistry.[12] This activity is conducted in groups of 2−3 students and proceeds in the order of 'Data Collection and Preprocessing', 'Exploratory Data Analysis", and "Modeling and Prediction" according to the inquiry worksheet (refer to "Inquiry Worksheet" in Supporting Information).

Through 'Data Collection and Preprocessing', students collect a dataset containing characteristics related to boiling point on various compounds. Students collect the dataset of compounds from PubChem. However, boiling points are not included in the PubChem dataset. Therefore, students aggregate data sets from two data sources. The first data source is PubChem and the second data source is CAS which provides the boiling point. Since the PubChem dataset contains the indiscriminate characteristics of compounds, the data preprocessing is necessary. In this study, data cleaning, data transformation, data discretization, and data integration methods are used in data preprocessing.

In "Exploratory Data Analysis", the preprocessed dataset is explored using Scatter plot and Correlations in Orange3′s visualization widget. Through the exploratory data analysis process, a new meaning can be discovered by examining the relationships of the characteristics of compounds.

The "Modeling and Prediction" is the process of creating classification and regression models through supervised ML. The preprocessed dataset is divided into training data and test data, and the model is trained using the training data. Once the model has been trained, the performance of each model is evaluated, and the best-trained optimal model is selected. This optimal model predicts the label and boiling point of arbitrary compounds that are not included in the PubChem and CAS dataset.

■ EXPERIMENT

This activity proceeds in the order of 'Data Collection and Preprocessing', 'Exploratory Data Analysis", and "Modeling and Prediction". (refer to "Inquiry Worksheet" in Supporting Information). Brief activities for each process are as followings:

Data Collection and Preprocessing

- Download the compound dataset from PubChem (refer to Table S1).
- Collect boiling point data from CAS.
- Delete unnecessary characteristics and compounds from the PubChem dataset.
- Add new characteristics such as numbers of C, N, O, and Side chain using the "LEN" and "SUBSTITUTE" functions in Excel (refer to "Inquiry Worksheet")
- Label compounds to "Hydrocarbon", "Alcohol", and "Amine" using the "IF" and "RIGHT" functions in Excel. (refer to "Inquiry Worksheet")
- Integrate boiling points from CAS to the preprocessed PubChem dataset. (refer to Table S2)

Exploratory Data Analysis

- Upload the preprocessed dataset (Table S2) to File widget of Orange3.
- Connect Correlations widget and Scatter plot widget with File widget.
- Analyze correlations and tendencies in dataset.

The workflow for data exploration is shown in Figure 1.

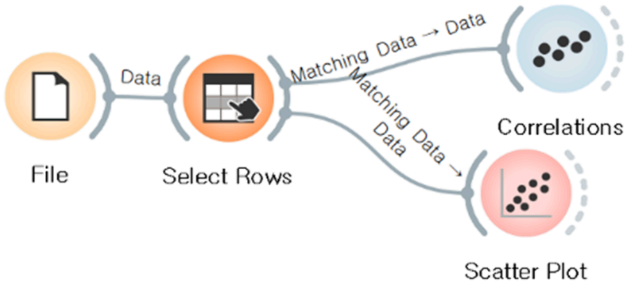


Figure 1. Workflow for the exploratory data analysis.

Modeling and Prediction

- Set the ratio of training data to test data in Data sampler widget.
- Select the optimal model for classification(kNN, Tree, Gradient Boosting, Random Forest, SVM, etc.) and regression models (Linear Regression, kNN, Tree, Random Forest, Gradient Boosting, AdaBoost, etc.).
- Predict labels or boiling points of the arbitrary compounds using the optimal model.

The workflow for modeling and prediction is shown in Figure 2.

Applying the Program to Students

In this activity, 20 students taking a general chemistry class participated in groups of two or three. This activity was conducted for 75 min a week for 4 weeks in class and was not related to grades. The instructor briefly demonstrated each

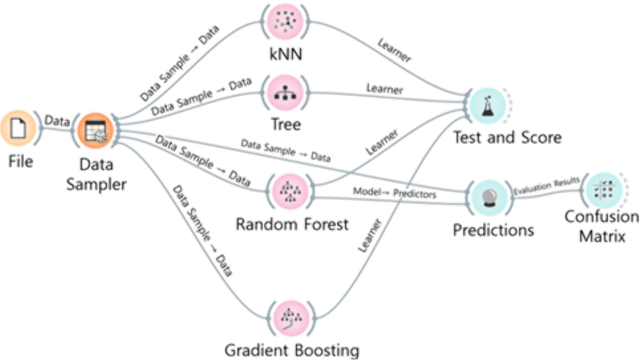
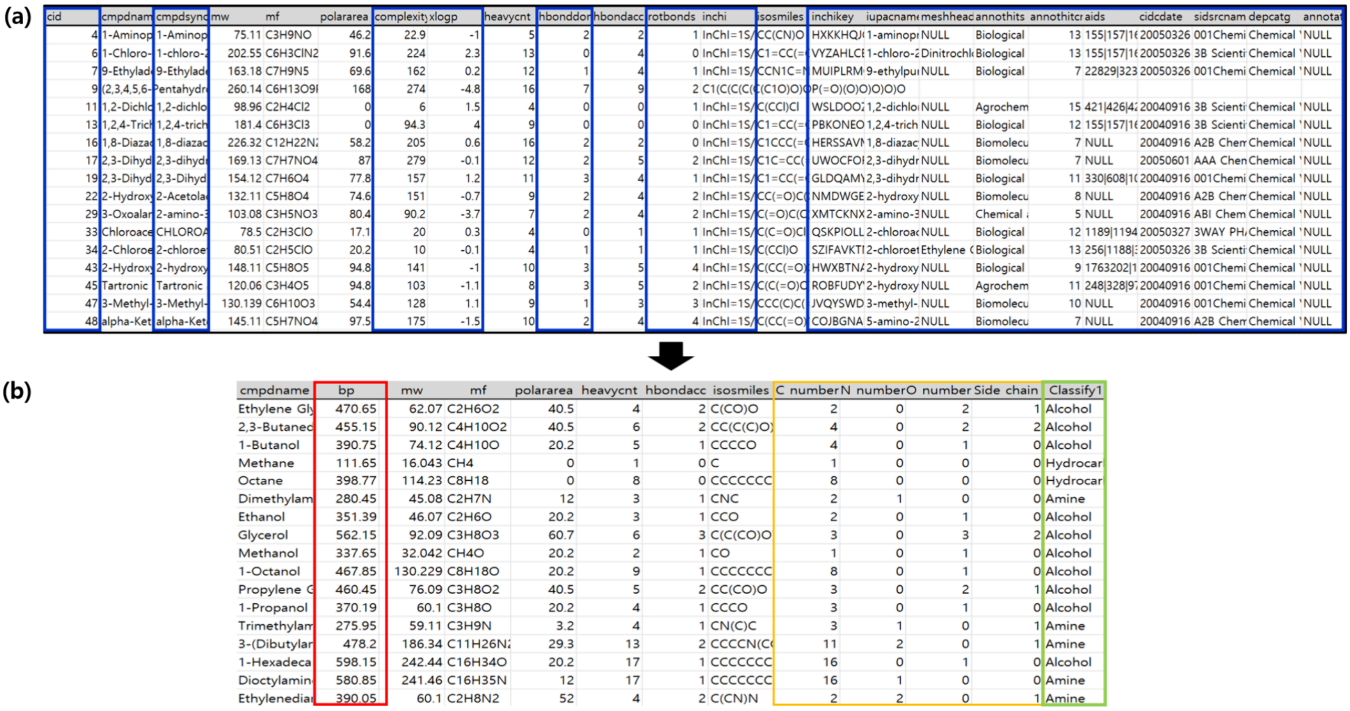


Figure 2. Workflow for modeling and prediction.

Table 1. Data Preprocessing From (a) Raw Dataset from PubChem to (b) Dataset with Data Preprocess Completed



(a)

cid	cmpdnam	cmpdsyno	mw	mf	polararea	complexity	xlogp	heavycnt	hbonddor	hbondacc	rotbonds	inchi	isosmiles	inchikey	iupacname	meshhead	annothits	annothitcr	aids	cidcdate	sidsrcnam	depcatg	annotat
4	1-Aminop	1-Aminop	75.11	C3H9NO	46.2	22.9	-1	5	2	2	1	InChI=1S/	CC(CN)O	HXKKHQJ	1-aminopr	NULL	Biological	13	155|157|1	20050326	001Chemi	Chemical '	NULL
6	1-Chloro-	1-chloro-2	202.55	C6H3ClN2	91.6	224	2.3	13	0	4	0	InChI=1S/	C1=CC(=	VYZAHLCE	1-chloro-2	Dinitrochl	Biological	13	155|157|1	20050326	3B Scienti	Chemical '	NULL
7	9-Ethylad	9-Ethylad	163.18	C7H9N5	69.6	162	0.2	12	1	4	1	InChI=1S/	CCN1C=N	MUIPLRM	9-ethylpu	NULL	Biological	7	22829|323	20050326	001Chemi	Chemical '	NULL
9	(2,3,4,5,6-	Pentahydr	260.14	C6H13O9P	168	274	-4.8	16	7	9	2	C1(C(C(C(C	(C1O)O)O	P(=O)(O)O)O)O)O									
11	1,2-Dichlo	1,2-dichlo	98.96	C2H4Cl2	0	6	1.5	4	0	0	1	InChI=1S/	C(CCl)Cl	WSLDOOZ	1,2-dichlo	NULL	Agrochem	15	421|426|4	20040916	3B Scienti	Chemical '	NULL
13	1,2,4-Trich	1,2,4-trich	181.4	C6H3Cl3	0	94.3	4	9	0	0	0	InChI=1S/	C1=CC(=	PBKONEO	1,2,4-trich	NULL	Biological	12	155|157|1	20040916	3B Scienti	Chemical '	NULL
16	1,8-Diazac	1,8-diazac	226.32	C12H22N2	58.2	205	0.6	16	2	2	0	InChI=1S/	C1CCC(=	HERSSAVI	1,8-diazac	NULL	Biomolecu	7	NULL	20040916	A2B Chem	Chemical '	NULL
17	2,3-Dihyd	2,3-dihydr	169.13	C7H7NO4	87	279	-0.1	12	2	5	2	InChI=1S/	C1C=CC(	UWOCFOI	2,3-dihydr	NULL	Biomolecu	7	NULL	20050601	AAA Chen	Chemical '	NULL
19	2,3-Dihyd	2,3-Dihyd	154.12	C7H6O4	77.8	157	1.2	11	3	4	1	InChI=1S/	C1=CC(=	GLDQAMY	2,3-dihydr	NULL	Biological	11	330|608|1	20040916	001Chemi	Chemical '	NULL
22	2-Hydroxy	2-Acetola	132.11	C5H8O4	74.6	151	-0.7	9	2	4	2	InChI=1S/	CC(=O)C(	NMDWGE	2-hydroxy	NULL	Biomolecu	8	NULL	20040916	A2B Chem	Chemical '	NULL
29	3-Oxoalar	2-amino-3	103.08	C3H5NO3	80.4	90.2	-3.7	7	2	4	2	InChI=1S/	C(=O)C(C	XMTCKNX	2-amino-3	NULL	Chemical a	5	NULL	20040916	ABI Chem	Chemical '	NULL
33	Chloroace	CHLOROA	78.5	C2H3ClO	17.1	20	0.3	4	0	1	1	InChI=1S/	C(C=O)Cl	QSKPIOLL	2-chloroa	NULL	Biological	12	1189|1194	20050327	3WAY PHA	Chemical '	NULL
34	2-Chloroe	2-chloroe	80.51	C2H5ClO	20.2	10	-0.1	4	1	1	1	InChI=1S/	C(CCl)O	SZIFAVKTI	2-chloroet	Ethylene (	Biological	13	256|1188|3	20050326	3B Scienti	Chemical '	NULL
43	2-Hydroxy	2-hydroxy	148.11	C5H8O5	94.8	141	-1	10	3	5	4	InChI=1S/	C(CC(=O)	HWXBTNA	2-hydroxy	NULL	Biological	9	1763202|1	20040916	001Chemi	Chemical '	NULL
45	Tartronic	Tartronic	120.06	C3H4O5	94.8	103	-1.1	8	3	5	2	InChI=1S/	C(C(=O)C	ROBFUDY'	2-hydroxy	NULL	Agrochem	11	248|328|9	20040916	001Chemi	Chemical '	NULL
47	3-Methyl-	3-Methyl-	130.139	C6H10O3	54.4	128	1.1	9	1	3	3	InChI=1S/	CCC(C)C(	JVQYSWD	3-methyl-	NULL	Biomolecu	10	NULL	20040916	001Chemi	Chemical '	NULL
48	alpha-Ket	alpha-Ket	145.11	C5H7NO4	97.5	175	-1.5	10	2	4	4	InChI=1S/	C(CC(=O)	COJBGNA	5-amino-2	NULL	Biomolecu	7	NULL	20040916	A2B Chem	Chemical '	NULL

(b)

cmpdname	bp	mw	mf	polararea	heavycnt	hbondacc	isosmiles	C number	N number	O number	Side chain	Classify1
Ethylene Gly	470.65	62.07	C2H6O2	40.5	4	2	C(CO)O	2	0	2	1	Alcohol
2,3-Butanec	455.15	90.12	C4H10O2	40.5	6	2	CC(C(C)O)	4	0	2	2	Alcohol
1-Butanol	390.75	74.12	C4H10O	20.2	5	1	CCCCO	4	0	1	0	Alcohol
Methane	111.65	16.043	CH4	0	1	0	C	1	0	0	0	Hydrocar
Octane	398.77	114.23	C8H18	0	8	0	CCCCCCC	8	0	0	0	Hydrocar
Dimethylam	280.45	45.08	C2H7N	12	3	1	CNC	2	1	0	0	Amine
Ethanol	351.39	46.07	C2H6O	20.2	3	1	CCO	2	0	1	0	Alcohol
Glycerol	562.15	92.09	C3H8O3	60.7	6	3	C(C(CO)O	3	0	3	2	Alcohol
Methanol	337.65	32.042	CH4O	20.2	2	1	CO	1	0	1	0	Alcohol
1-Octanol	467.85	130.229	C8H18O	20.2	9	1	CCCCCCC	8	0	1	0	Alcohol
Propylene G	460.45	76.09	C3H8O2	40.5	5	2	CC(CO)O	3	0	2	1	Alcohol
1-Propanol	370.19	60.1	C3H8O	20.2	4	1	CCCO	3	0	1	0	Alcohol
Trimethylam	275.95	59.11	C3H9N	3.2	4	1	CN(C)C	3	1	0	1	Amine
3-(Dibutylar	478.2	186.34	C11H26N2	29.3	13	2	CCCCN(C(	11	2	0	1	Amine
1-Hexadeca	598.15	242.44	C16H34O	20.2	17	1	CCCCCCC	16	0	1	0	Alcohol
Dioctylamin	580.85	241.46	C16H35N	12	17	1	CCCCCCC	16	1	0	0	Amine
Ethylenediar	390.05	60.1	C2H8N2	52	4	2	C(CN)N	2	2	0	1	Amine

process, and students conducted inquiry activities using their laptops. Students installed Orange3 at home by themselves, according to the manual. Since they had no experience using Orange3 and ML, students practiced ML with Orange3 using well-known iris and abalone data.[13,14] After the practice, students performed the inquiry activities of 'Data Collection and Preprocessing', 'Exploratory Data Analysis", and "Modeling and Prediction". The students had difficulty with AI/DS activities because they had no experience with them. The instructor intervened only when students were having difficulty. Most students understood each process well, and the group activities were successfully completed. The degree of goal achievement was confirmed by checking the inquiry results recorded in the worksheets.

A postsurvey on 'AI education satisfaction' and 'data literacy' was conducted on 20 students on a Likert scale, with 5 meaning strongly agree. The survey of AI education satisfaction is divided into the "AI education confidence" and "AI education satisfaction". The survey of data literacy is divided into the 'Data Collection and Preprocessing", "Data Analysis", and 'Data Prediction and Evaluation'. This survey consists of a total of 18 questions and detailed questions are specified in Table S3.

■ DISCUSSION

The learning goal of this activity is for students to experience the application of machine learning to chemistry content and to recognize the importance of data science in machine learning.

The preprocessing of raw dataset as ML input dataset, the exploratory data analysis, and classifying compounds and predicting boiling points would be described.

Preprocessing of Raw Dataset as ML Input Dataset

The raw dataset collected from PubChem is shown in the following Table 1(a). This dataset contains 24 characteristics (cid, cmpdname, cmpdsynonym, mw, mf, polararea, heavycnt, hbondacc, isosmiles, etc.) for 277,569 compounds but no data for boiling point. The data preprocess, including data cleaning, data transformation, and data discretization, was performed. The data cleaning process leaves only necessary characteristics-(cmpdname, mw, mf, polararea, heavycnt, hbondacc, isosmiles) and necessary compounds(hydrocarbon, alcohol, amine). Through the data transformation process, 4 characteristics (C/N/O number and side chain number) were added. With data cleaning and transformation completed, a data discretization process (labeling compounds as hydrocarbons, alcohols, and amines) was performed for exploratory data analysis and classification modeling. The blue boxes in Table 1(a) were deleted through the data cleaning process. In Table 1(b), the yellow box goes through the data transformation process and the green box through the data discretization process. As a result of data discretization, the dataset contains 12 characteristics for a total of 4,717 compounds. And boiling points were merged into this dataset through a data integration process using Colab. This is expressed in the red box in Table 1(b). After deleting compounds without boiling point, dataset with 13 characteristics for a total of 1,748 compounds remains.

Exploratory Data Analysis

Using the Correlations widget, the results of examining the correlation coefficients of boiling point (bp), molecular weight(mw), number of non-hydrogen atoms (heavycnt), area of the polar area of the molecule (polararea), and number of hydrogen bond acceptors (hbondacc) are shown in the Heatmap (Table2). Through this, it could be confirmed that there is a very high correlation between bp and mw, bp and Heavycnt, Heavycnt and mw, and Polararea and Hbondacc. Looking at the correlation coefficient with boiling point, the correlation coefficient between boiling point and molecular weight was the highest at +0.872.

Table 2. Correlation Coefficients among bp, mw, heavycnt, polararea, and hbondacc

	bp	mw	heavycnt	polararea	hbondacc
bp		+0.872	+0.854	+0.210	+0.204
mw	+0.872		+0.996	−0.0082	−0.0068
heavycnt	+0.854	+0.996		−0.136	−0.123
polararea	+0.210	−0.0082	−0.136		+0.981
hbondacc	+0.204	−0.0068	−0.123	+0.981	

The bp tendency of mw and heavycnt is shown in the Figure 3. In Figure 3(a), it can be seen that the boiling point generally increases as the molecular weight increases. In addition, the boiling point was generally higher in the order of alcohol > amine > hydrocarbon with similar molecular weights.[15] To clearly compare the boiling point tendency among alcohols, amines, and hydrocarbons, the relationship between bp and heavy compound was examined as shown in Figure 3(b). As a result, the boiling point increases in the order of alcohol, amine, and hydrocarbon when the number of non-hydrogen atoms is the same.

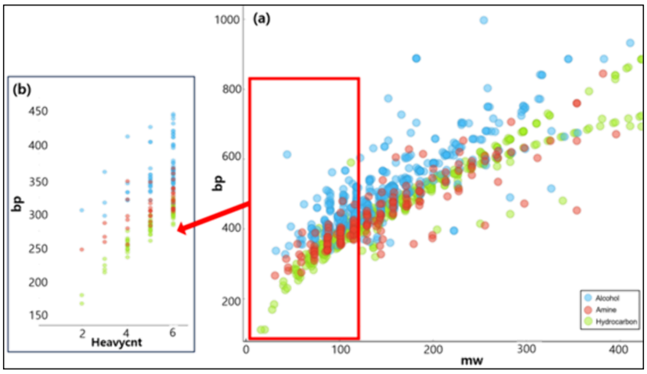
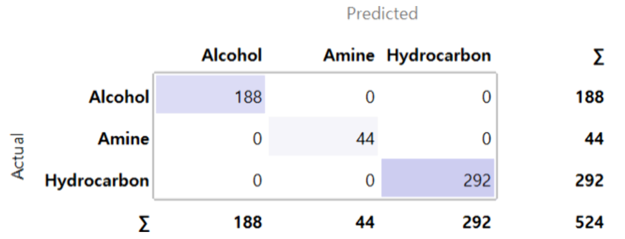


Figure 3. Exploration data analysis process (visualization): (a) bp-mw and (b) bp-heavycnt.

Classification Model for Label Prediction

This process is intended to select the optimal classification model and predict the label of compounds. To classify compounds, machine learning algorithms such as kNN, Tree, Random Forest, and Gradient Boosting were learned using training data.[16] Random Forest was selected for evaluation. Random Forest, an ensemble learning method, works by constructing multiple decision trees during the training phase and outputting the average prediction of the individual trees. Its advanced capabilities allowed for a more nuanced and accurate prediction of boiling points, demonstrating the significant advantages of machine learning models over traditional methods in chemical data analysis.[16] The prediction results of Random Forest using the test data are shown in Table 3. The labels of 524 test data, 30% of the 1748 data, were accurately predicted.

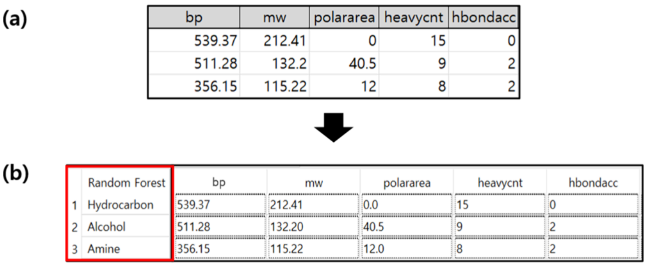
Table 3. Predict Results of Classification Model by the Confusion Matrix Widget



		Predicted			
		Alcohol	Amine	Hydrocarbon	Σ
Actual	Alcohol	188	0	0	188
	Amine	0	44	0	44
	Hydrocarbon	0	0	292	292
	Σ	188	44	292	524

Since all test data were accurately predicted, the Random Forest model was used to predict labels of an unknown dataset. The unknown dataset has 12 characteristics without label of compounds. Only numerical characteristics (bp, mw, polararea, heavcnt, hbondacc) were used when the model predicted labels. Compounds of unknown dataset consisted of 6-Methyltetradecane, 1,5-heptanediol, and 1-butanamine. As a result of the prediction, it was confirmed that 6-Methyltetradecane was accurately predicted as a hydrocarbon, 1,5-heptanediol as an alcohol, and 1-butanamine as an amine (Table 4).

Table 4. Prediction Results of Classification Model for Unknown Dataset. (a) Input Data. (b) Output Result



(a)

bp	mw	polararea	heavycnt	hbondacc
539.37	212.41	0	15	0
511.28	132.2	40.5	9	2
356.15	115.22	12	8	2

(b)

	Random Forest	bp	mw	polararea	heavycnt	hbondacc
1	Hydrocarbon	539.37	212.41	0.0	15	0
2	Alcohol	511.28	132.20	40.5	9	2
3	Amine	356.15	115.22	12.0	8	2

Regression Model for Boiling Point Prediction

This process is intended to select the optimal regression model and predict the boiling point of compounds. To predict boiling point, machine learning algorithms such as kNN, Tree, Random Forest, Linear Regression were learned using training data.[16] Random Forest model had the highest performance based on the Mean Squared Error (MSE, 1874.417), square root of MSE (RMSE, 43.295), Mean Absolute Error (MAE, 19.876), and R-squared (R2, 0.881) indices. It was confirmed that the selected optimal model, Random Forest, predicted the test data well. Since test data were well predicted, Random Forest model was used to predict boiling point of unknown dataset. The unknown dataset consisted of 5 compounds: 2 hydrocarbons(3-ethyl-5-methylheptane, Octatetracontane), 2 alcohols(1-Pentacosanol, 3-pentyn-1-ol), and 1 amine(1,2,2-Trimethylpropylamine). The unknown dataset has 12 characteristics without boiling point. As a result, the five compounds showed errors of 0.83, 7.12, 3.9, 9.94, and 0.65 compared to the theoretical value,[17] respectively. Since these predicted values are all within the error range of the theoretical value, this random forest model predicted the boiling point of the compound well. The prediction results are within the average error range of 4.49. Prediction results of regression model for unknown dataset are shown in Table 5.

Results of Applying the Program to Students

A postsurvey on “AI education satisfaction” and “data literacy” was conducted on 20 students using a 5-point Likert scale, with 5 meaning strongly agree. The average score of “AI education satisfaction” was 3.96 point, and the average score of “data literacy” was 3.77 point. As such, there are generally positive responses in both areas.

In the ’AI education satisfaction’ survey, which consists of ’AI education confidence’ and ’AI education satisfaction’, the

Table 5. Prediction Results of Regression Model for Unknown Dataset

Compound	Predicted value [K]	Theoretical value [K]	Error
3-ethyl-5-methylheptane	432.32	433.15 ± 7	0.83
Octatetracontane	848.97	841.85 ± 13	7.12
1-Pentacosanol	672.65	676.55 ± 8	3.9
3-pentyn-1-ol	419.71	429.65 ± 13	9.94
1,2,2-Trimethylpropylamine	371.85	372.45 ± 8	0.65

perception of ’AI education confidence’(3.83) was found to be slightly lower than the perception of ’AI education satisfaction’(4.06). This can be interpreted as the fact that although the class content was interesting, there were some difficulties in the class.

The survey of “data literacy” is divided into ‘Data Collection and Preprocessing”, “Data Analysis”, and ‘Data Prediction and Evaluation’.

Among these, the literacy of ‘Data Prediction and Evaluation’(3.86) had the highest average score than the literacy of ‘Data Collection and Preprocessing’ (3.61) and “Data Analysis”(3.83), because students were able to more easily predict the labels and boiling points of compounds using NCLC Orange3. On the other hand, literacy of ‘Data Prediction and Evaluation’ had the lowest score than other literacies because students felt that the process of collecting and preprocessing data was much longer and more complicated than other areas.

CONCLUSIONS

“Education the bilinguals of the future” means training people who can create new knowledge and value by integrating AI and DS into their field of expertise. In this respect, it is necessary to develop materials and apply classes that allow students to experience data processing and the use of AI in scientific topics. In this study, first of all, students experience collecting and preprocessing a dataset consisting of various characteristics that can affect boiling point. There are some well-refined data sets available for training ML. However, in order to actually run ML on the field of interest, most data requires a preprocessing. Although NCLC ML tools have been developed and deployed,[18,19] students must preprocess the raw data sets themselves. This boiling point learning material, which provides experience in the process of deleting unnecessary data, adding more necessary data, and regenerating some data, will have great implications for convergence education. This may be the reason why students answered in the survey that the preprocessing process was relatively difficult.

Second, students experience exploratory data analysis on the compound dataset. Exploratory data analysis helps students find correlations among characteristics of the data. After finding correlations, a scientific and logical explanation process is needed to explain the relationship of characteristics. This process may greatly contribute to finding new values. On the boiling points of compounds, students found that boiling point was highly correlated to molecular weight or C/N/O number. The existence of this correlation does not necessarily explain that there is a causal relationship between the boiling point and molecular weight or C/N/O number. In other words, correlations between characteristics must be combined with expert knowledge in the field to better explain the inter-relationships between characteristics.

Third, students experience operating ML on the dataset to classify compounds and predict boiling points. ML uses well-refined data to create a model that can explain data patterns and predict unknown characteristics. Students create two models; classification model for label prediction and regression model for boiling point prediction. Using the created models, students could predict label and boiling point of the compound from unknown data sets. In this way, students developed the ability to apply machine learning in various fields through the experience of creating models and predicting characteristics. Through this, students could learn how to gain insight from the data. They will be able to gain new insights.

In summary, as the need for AI-subject convergence has recently increased in the field of education, this activity introduces the convergence of chemistry and machine learning, including data preprocessing. This activity provides examples of how AI and DS technologies can be integrated into each domain. These convergence activities will help to improve conceptual understanding and AI literacy through visualization of data and the modeling process. This activity will provide meaningful implications for research exploring learning methods that efficiently construct knowledge in each subject by combining AI and DS technologies.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available at https://pubs.acs.org/doi/10.1021/acs.jchemed.3c01040.

Table S1 (ZIP)

Table S2 (XLSX)

Table S3. Contents of survey questions according to survey area (PDF)

Table S3. Contents of survey questions according to survey area (DOCX)

Inquiry Worksheet (PDF)

Inquiry Worksheet (DOCX)

AUTHOR INFORMATION

Corresponding Author

Seong-Joo Kang − *Department of Chemistry Education, Korea National University of Education, Cheongju, Chungbuk 28173, Republic of Korea;* orcid.org/0000-0002-1531-1704; Email: sjkang@knue.ac.kr

Authors

Shin-Yu Kim − *Department of Chemistry Education, Korea National University of Education, Cheongju, Chungbuk 28173, Republic of Korea*

Inseong Jeon − *Department of Computer Education, Korea National University of Education, Cheongju, Chungbuk 28173, Republic of Korea*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jchemed.3c01040

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Vial, G. Understanding digital transformation: A review and a research agenda. *Managing Digital Transformation* **2021**, 13−66.

(2) Jakobsche, C. E.; Kongsomjit, P.; Milson, C.; Wang, W.; Ngan, C. K. Incorporating an Intelligent Tutoring System into the DiscoverOChem Learning Platform. *Journal of Chemical Education* **2023**, 3081−3088, DOI: 10.1021/acs.jchemed.3c00117.

(3) PINTO DOS SANTOS, D.; BAEßLER, B. Big data, artificial intelligence, and structured reporting. *European radiology experimental* **2018**, *2* (1), 42.

(4) Lohr, S. Plans College for Artificial Intelligence, Backed by $1 Billion. *New York Times*, October 15 2018

(5) Dans, E. MIT Knows That AI Is The Future. *Forbes.*, Oct. 16, **2018**.

(6) OH, P. K.; KANG, S. J. Integrating Artificial Intelligence to Chemistry Experiment: Carbon Dioxide Fountain. *J. Chem. Educ.* **2021**, *98* (7), 2376−2380.

(7) SUFI, F. Algorithms in low-code-no-code for research applications: a practical review. *Algorithms* **2023**, *16* (2), 108.

(8) Villegas-Ch, W.; García-Ortiz, J.; Sánchez-Viteri, S. Identification of the factors that influence university learning with low-code/no-code artificial intelligence techniques. *Electronics* **2021**, *10* (10), 1192.

(9) Beranic, T.; Rek, P.; Heričko, M. Adoption and usability of low-code/no-code development tools. *Central European Conference on Information and Intelligent Systems* **2020**, 97−103.

(10) Orange3. https://orangedatamining.com/ (accessed October 18, 2023).

(11) Joss, L.; Müller, E. A. Machine learning for fluid property correlations: classroom examples with MATLAB. *J. Chem. Educ.* **2019**, *96* (4), 697−703.

(12) Reyes, R. L. Exploring Science Literature: Integrating Chemistry Research with Chemical Education. *J. Chem. Educ.* **2023**, *100* (6), 2303−2311.

(13) *Kaggle*. https://www.kaggle.com/datasets/uciml/iris (accessed December 13, 2023).

(14) Kaggle. https://www.kaggle.com/datasets/rodolfomendes/abalone-dataset (accessed December 13, 2023).

(15) MURPHY, P. M. Teaching structure−property relationships: Investigating molecular structure and boiling point. *J. Chem. Educ.* **2007**, *84* (1), 97.

(16) Mahesh, B. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)* **2020**, *9* (1), 381−386.

(17) Chemspider. https://www.chemspider.com/ (accessed December 6, 2023).

(18) Knack. https://www.knack.com/ (accessed September 30, 2023).

(19) *Googleappsheet*. https://about.appsheet.com/home/ (accessed September 30, 2023).