JOURNAL OF CHEMICAL EDUCATION

pubs.acs.org/jchemeduc Article

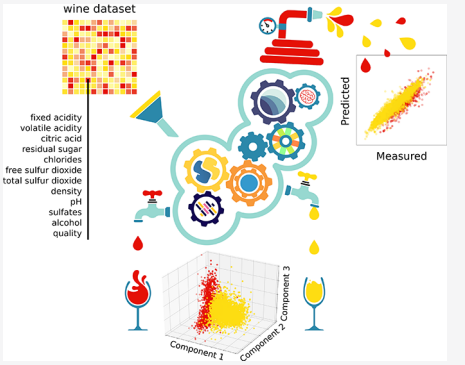
A Gentle Introduction to Machine Learning for Chemists: An Undergraduate Workshop Using Python Notebooks for Visualization, Data Processing, Analysis, and Modeling

Deborah Lafuente, Brenda Cohen, Guillermo Fiorini, Agustín Alejo García, Mauro Bringas, Ezequiel Morzan, and Diego Onna*

Cite This: https://doi.org/10.1021/acs.jchemed.1c00142 Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Machine learning, a subdomain of artificial intelligence, is a widespread technology that is molding how chemists interact with data. Therefore, it is a relevant skill to incorporate into the toolbox of any chemistry student. This work presents a workshop that introduces machine learning for chemistry students based on a set of Python notebooks and assignments. Python, one of the most popular programming languages, is open source, free to use, and has plenty of learning resources. The workshop is designed for students without previous experience in programming, and it aims for a deeper understanding of the complexity of concepts in programming and machine learning. The examples used correspond to real data from physicochemical characterizations of wine, a content that is of interest for students. The contents of the workshop are introduction to Python, basic statistics, data visualization, and dimension reduction, classification, and regression.



KEYWORDS: *Upper-Division Undergraduate, Chemoinformatics, Interdisciplinary/Multidisciplinary, Computer-Based Learning, Chemometrics, Computational Chemistry*

INTRODUCTION

The amount and complexity of data grows rapidly every year.[1,2] Therefore, analyzing this data with classical methods results in a more challenging and impractical task. Data analysis and scientific computing are becoming an important part of laboratory procedures following the advances in automation and the Internet of things (IoT), which generates more relevant data from chemical systems as several physicochemical variables are measured.[3] These changes in data analysis trigger the need for new generations of chemistry students to learn modern tools, requiring the development of novel skills that are not yet covered in most chemistry undergraduate programs.[4] In this line, the integration of huge amounts of data and artificial intelligence is regarded as the "fourth paradigm of science",[5] and the number of possible applications in the chemical field is rising notably. Machine learning is a subdomain of artificial intelligence that has evolved especially in recent years.[6,7] It allows scientists with enough data and a proper algorithm to discover rules that can be used to solve problems by building predicting models whose performance can be assessed. The data is labeled when the output are categories or values associated with the features (input or variables). The process of learning to solve the problem with labeled data is known as supervised learning. The relation of features, model, and labels are shown in eq 1. Unsupervised learning indicates that the training process is performed on data without labels.

$$\text{Features} \rightarrow \text{model} \rightarrow \text{Labels} \quad (1)$$

In the classroom, computational methods are usually applied for chemical system simulations, such as DFT calculations or molecular dynamics simulations to predict or interpret experimental results, or for data analysis from experimental results. Experiences like these have been published in previous issues of this journal.[8−15] There are also two remarkable courses for teaching computational skills to chemistry students by Weiss[16,17] and Menke.[18] Both of them cover topics such as an introduction to programming with Python, performing simple simulations, and analyzing data from tables, spectra, and images. Another relevant course for cheminformatics was organized by the Committee on Computers in Chemical Education, a part of the American Chemical Society's Division of Chemical Education.[19] Regarding machine learning, the

Received: February 12, 2021
Revised: July 25, 2021

material developed for teaching chemistry students is more scarce. An in-depth discussion of the methodologies for chemical systems can be found elsewhere.[20] For example, Joss et al. presented a classroom activity for predicting the normal boiling point of organic compounds using multivariate linear regression and artificial neural networks, two typical machine learning algorithms.[21] In turn, Antonelli et al. generated a lab activity for multivariate calibration using partial least-squares regression on the programing language R.[22] These hands-on experiences are a valuable asset for incorporating computational skills via the exposition of new tools to solve relevant problems in chemistry and, by doing so, to promote collaborative work and the democratization of knowledge.[13]

Several general courses consisting of one or two sessions have been implemented before, including programming, data visualization, and machine learning, among others topics. As an example, *The Carpentries*,[23] a project focused on teaching computing skills to researchers, has hosted a number of such courses. Another important organization is the *Molecular Sciences Software Institute* (MolSSI), which has organized several computational chemistry-related courses.[24]

In this work, we present a workshop for introducing chemistry students to machine learning, which was designed to offer some essential skills for all chemists such as processing, analysis, visualization, and modeling large, complex, and multidimensional data. The workshop is based on five Python notebooks and their corresponding assignments to guide the students from learning Python language to developing machine learning models. Along this journey, several machine learning skills and concepts are covered for students to get some degree of familiarity with basic statistics, data visualization, and dimension reduction, classification, and regression. The chosen data set is a physicochemical characterization of red and white wine, a topic both interesting for students and relevant as a case of study. This workshop was organized in two syncronic 3 h online sessions, lasting 2 days. In addition to this, the assignments demanded around 15 hours to complete.

OBJECTIVES

This workshop aims to expose chemistry students to machine learning, including some programming notions, and data visualization, processing, analysis, and modeling. It is expected that, by the end of the workshop, the students are able to

1. write and understand simple code using Python;
2. implement a machine learning algorithm for analyzing a wine data set using several scientific Python libraries;
3. develop a self-learning attitude through the use of documentation and question and answer (Q&A) sites for programmers.

PROGRAMMING LANGUAGE AND SOFTWARE DETAILS

This workshop exposes students to tools for data analysis and is based on Python 3 and its scientific libraries. Python is an object-oriented programming language that is widely used for a variety of purposes, which runs on multiple platforms. Python is also free and open-source, making it accessible to all students, and it has a simple and straightforward syntax, which makes it a convenient choice for students with no prior programming experience. There are also plenty of online resources for Python, and a large community of users, that allows students to solve minor programming issues with a simple web search when coding unsupervised. Lastly, there are a lot of developer tools for Python, as well as free or accessible online Python courses for students willing to expand their knowledge.

Google Colab notebooks is a web application that allows users to create and share documents that contain live code, visualizations, and narrative text.[25] It was used in this workshop because explanatory text, images, formulas, and executable Python code can be combined in one single document. By means of Google Colab notebooks, the user can write and execute Python code from their browser with no installation required, and it runs the code in Google Cloud servers, using Google's processing units regardless of the user's hardware limitations. Additionally, notebooks can be executed using a smartphone, which improves accessibility in developing countries.[26,27]

Python offers a wide variety of libraries for data analysis and machine learning that can be applied to the processing of chemical data and chemical research. The use of libraries allows for faster and more succinct code, leaving the user more time to focus on the data analysis rather than on the code's complexity. Furthermore, Python's most popular libraries are open-source. The libraries used along the workshop are NumPy, Pandas, Seaborn, Matplotlib, Pingouin, and scikit-learn, as shown in Table 1. NumPy offers the possibility of working with arrays of numbers, as well as a variety of functions for working in the domain of linear algebra and matrices. Pandas allows for the treatment of large data sets in the convenient structure of dataframes and various tools for their handling. A dataframe is a two-dimensional mutable data structure with rows and columns that simplifies data manipulation. Matplotlib is a *de facto* library for creating static, animated, and interactive visualizations. Seaborn, together with Matplotlib, simplifies the task of creating intricate graphs by writing a few lines of code. This comes in good use when plotting visualizations for preliminary data analysis. Seaborn also offers some integrated example data sets that were used in one of the notebooks and one of the assignments. Pingouin is a statistical package that allows the performance of simple analysis on Pandas data structures. Plotly is an advanced plotting library that allows the creation of interactive plots easily. Scikit-Learn has various tools for preprocessing data that are effective and simple to use. It also offers a series of machine learning functions, such as classification, regression, and clustering algorithms. It is built on NumPy, SciPy, and

Table 1. Python Libraries Used in the Notebooks to Analyze and Plot the Data

library name	description	web site[a]
Matplotlib	plotting tools	https://matplotlib.org/
NumPy	linear algebra and fast math library	https://numpy.org/
Pandas	handling and preprocessing of large data sets	https://pandas.pydata.org/
Pingouin	common statistical tests	https://pingouin-stats.org/
Seaborn	advanced plotting functions	https://seaborn.pydata.org/
Plotly	interactive plots and advanced plotting functions	https://plotly.com/python/
Scikit-Learn	scientific computing and machine learning tools	https://scikit-learn.org/

[a]Web sites accessed July 2021.

https://doi.org/10.1021/acs.jchemed.1c00142
J. Chem. Educ. XXXX, XXX, XXX−XXX

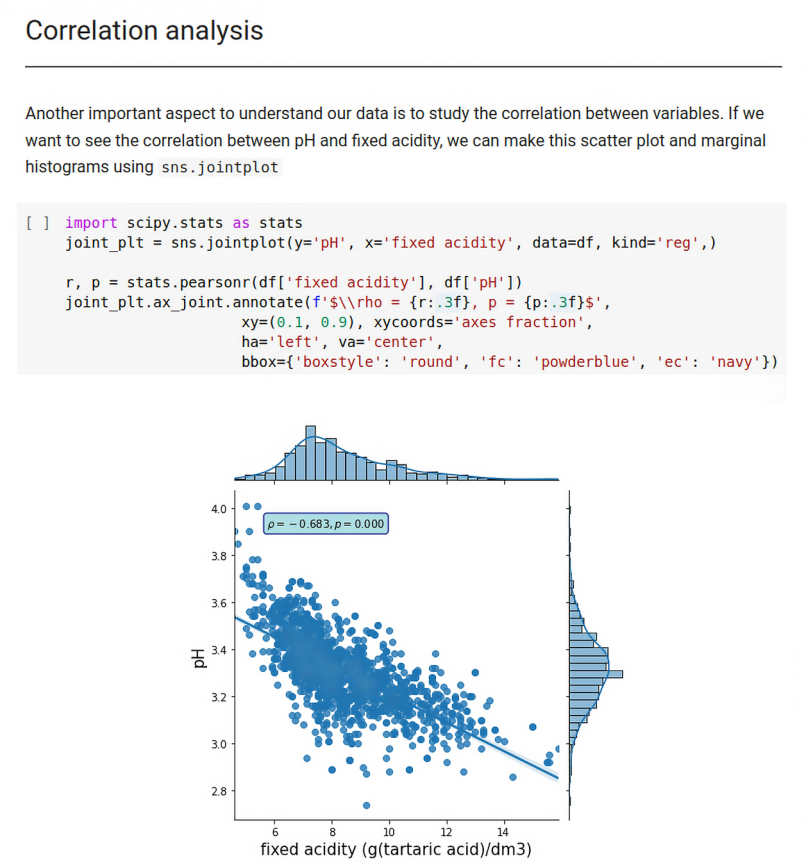


Figure 1. Screenshot of a notebook containing code (light gray section), text for explanations (white section), and a plot.

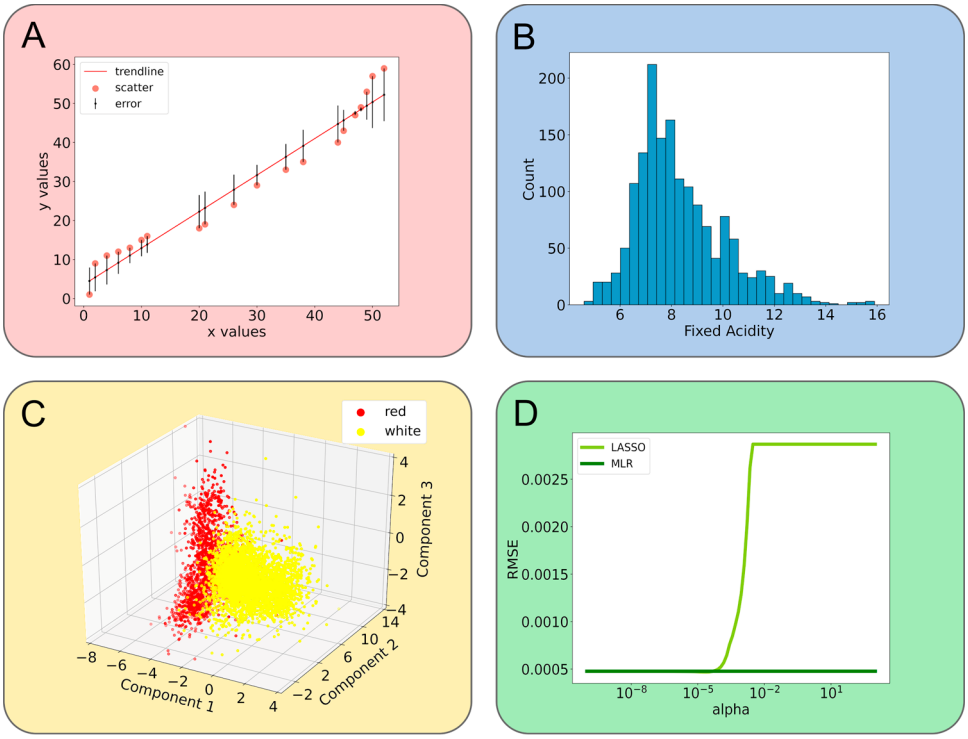


Figure 2. Different types of plots were generated throughout the notebooks and assignments. (A) Linear regression with its error bars from a scatter plot. (B) Histogram of fixed acidity values. (C) A 3D plot of principal component analysis from red and white wine. (D) Root mean square error (RSME) on the prediction of density for wine using multiple linear regression (MLR) and least absolute shrinkage and selection operator (LASSO) at different penalization of the coefficient (alpha values). RSME of MLR is constant as it is independent of alpha.

https://doi.org/10.1021/acs.jchemed.1c00142
J. Chem. Educ. XXXX, XXX, XXX−XXX

Matplotlib, so a vast number of visualizations can be plotted after applying the mentioned algorithms to the data, which can come to good use in preliminary data analysis.

WORKSHOP STRUCTURE AND CONTENT

This workshop is designed as a stand-alone workshop as most curricula do not yet include machine learning. Nevertheless, it could be easily included as a module in analytical or instrumental chemistry courses. The proposed workshop is divided into five Python notebooks created for students without previous programming or Python knowledge. It first introduces simple Python statements, and its complexity increases as it incorporates programming and machine learning concepts. Also, five assignments with key are included. All files are in jupyter notebooks' format to combine code with explanations written in markdown format, and plots for making a flexible and potent teaching tool, as shown in Figure 1. The notebooks are available[28] at GitHub both in English and Spanish and are described below.

Notebook 1: Introduction to Python

The first notebook focuses on introducing students to Python from a practical point of view, mainly teaching them how to use functions and advanced libraries. It covers variables, data types, value assignment, and value comparison. Pandas is presented for data handling and pandas dataframes and Seaborn is applied for fitting and plotting a linear regression.

Assignment 1

This assignment is designed to reinforce the concept of functions and libraries by asking students to (i) program a calculator, (ii) work with lists, strings, and become familiar with data types, (iii) plot data from a dataframe using Seaborn, (iv) apply NumPy functions, (v) perform a linear regression with NumPy, and (vi) plot a graph with Matplotlib (Figure 2A).

Notebook 2: Basic Statistics

Here, different ways for importing data are covered and the wine data set is presented. This data set is the main data analyzed for the rest of the workshop. After loading the data in a dataframe, basic methods for Pandas dataframes are explored (head, min, max, mean, std, describe). A confidence interval is calculated as an introduction to errors and statistics. Histograms (Figure 2B), box plots, and violin plots are examined using Seaborn. Finally, hypothesis tests (one-sample and two-unpaired-samples two-tailed t tests) and analysis of variance (ANOVA) are performed using Pingouin.

Assignment 2

Histograms, boxplots, confidence intervals, and ANOVA tests are implemented. More advanced contents include the exploration of pseudorandom number generation, the creation of a grid of histograms using Pandas, and the examination of binomial, poisson, and normal distributions. Also, the central limit theorem is exhibited for various distributions due to its implications for data standardization.

Notebook 3: Visualization and Dimension Reduction

Visualization is key in understanding complex data sets with several variables. Accordingly, box plots and violin plots are used for representing multivariable data. Another important tool to explore data is correlation analysis, as several methods rely on a dependency between variables. The Pearson correlation coefficient is calculated using scipy, and it is represented with heatmaps and pair plots using Seaborn. After this, dimension reduction is explored for high-dimensional data. Principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP) methods are implemented. Finally, advanced plots such as interactive plots with ipywidgets and 3D plots with Matplotlib and Plotly (Figure 2C) are also covered.

Assignment 3

Box plots are used for inspecting multiple variables simultaneously, and heatmaps and scatter plots for analyzing the correlation between variables (Pearson correlation coefficient). To clarify dimension reduction, PCA is calculated step by step using algebraic equations instead of Scikit-Learn functions.

Notebook 4: Classification

The task of classifying implies dividing data among classes according to predictor variables. In this case, the challenge is to discriminate between red and white wine using physicochemical attributes. The classification methods covered are logistic regression and decision tree. The concepts of data splitting, data standardization, confusion matrix, and metrics calculation (accuracy, precision, and recall) are analyzed in this instance.

Assignment 4

Test-train splitting is performed on the wine data set. K-nearest neighbors classification is applied and performance metrics are calculated. The meaning and implications of the value of K are analyzed in terms of accuracy.

Notebook 5: Regression

Regression analysis is performed to predict wine density based on other physicochemical properties. Correlations between other attributes and density are calculated. Both simple and multiple linear regression analyses are performed, and the importance of standardization is presented. Lastly, the LASSO regularization strategy is applied, paying attention to the magnitude and meaning of the regularization constant (Figure 2D).

Assignment 5

Following the density prediction problem, a second regularization strategy (ridge) is proposed. Once again, the magnitude of the regularization constant is explored and a discussion of the differences between L1 and L2 regularization methods is opened.

DATASET DESCRIPTION

In this workshop, the data set explored is a collection of white and red wine from the "Vinho Verde" wine region in Portugal.[29] The data set for red wine consists of 1599 samples and the one for white wine consists of 4898 samples. Each sample has 12 physicochemical variables as shown in Table 2. This data set is free and available in a UC Irvine machine learning repository.[30]

ONLINE DISCUSSION AND TUTORING PLATFORM

During and after the workshop, communication was managed using Discord, a platform designed to communicate among video game players.[31] Different communities use it for hobbies, application development, among others, giving it a rising popularity. Discord allows communication via voice, video, and text chat, video calling, and screen sharing. Also, files and links

https://doi.org/10.1021/acs.jchemed.1c00142
J. Chem. Educ. XXXX, XXX, XXX−XXX

Table 2. Dataset Description from Wine Dataset Containing Sample Quantities of Each Wine Type, Wine Physicochemical Attributes, and Their Corresponding Units

input variables		
no.	features	units
1	fixed acidity	g(tartaric acid)/dm^3
2	volatile acidity	g(acetic acid)/dm^3
3	citric acid	g/dm^3
4	residual sugar	g/dm^3
5	chlorides	g(sodium chloride)/dm^3
6	free sulfur dioxide	mg/dm^3
7	total sulfur dioxide	mg/dm^3
8	density	g/cm^3
9	pH	-
10	sulfates	g(potassium sulfate)/dm^3
11	alcohol	vol %

output variable		
nos.	labels	units
12	quality	-

samples	type
1599	red wine
4898	white wine

can be shared on group messages (channels) or private messages. This is an adequate communication platform for creating a community of students, former students, and faculty members, creating an open and horizontal environment for continuous updating, collaboration, and networking in topics related to machine learning in chemistry. It is worth mentioning that this workshop was entirely online due to the social isolation during the COVID-19 pandemic and it did not present any complications. Therefore, it could be an excellent complement for laboratory experiments at home.[32]

STUDENTS PREREQUISITES

This workshop requires some knowledge of statistics and analytical chemistry, and it could therefore be implemented in any undergraduate chemistry program after the second year. The material was not intended as deep explanations of the different topics, as the authors consider that each of them could be covered in a standalone workshop. On the contrary, the material is an introduction to machine learning that, at the same time, constitutes the base for data analysis in chemistry, a necessary skill to comply with the industry and research advances in automation and the Internet of things.[33]

In order to minimize students executing commands without understanding the reason for it, they need some basic skills like procedural thinking, plotting, and data interpretation, which imply some knowledge of data analysis software like MS excel or Google sheets. Also, students should be comfortable dealing with large amounts of data when it becomes impossible to check every data point and calculation manually.

WORKSHOP IMPLEMENTATION

This workshop was structured in two online sessions of 3 h each, and the assignments demanded another 15 hours. The sessions were recorded allowing students to revisit the content as many times as necessary. Also, a Discord server was established as a Q&A forum for students to submit questions or queries and get replies about workshop-related issues. Approximately 150 students enrolled to attend this workshop, which is an extracurricular activity as machine learning is not yet included in the curricula. Approximately 100 students attended the first synchronic session and 65 attended the second. Even though it was announced as an undergraduate workshop, graduate, Ph.D. candidates, postdocs, and young independent researchers also enrolled.

During the design of the workshop, it was assumed that students would have little to no experience in programming in Python; this was later confirmed by the survey students completed upon enrollment (Figure 3A), as 58.2% of students

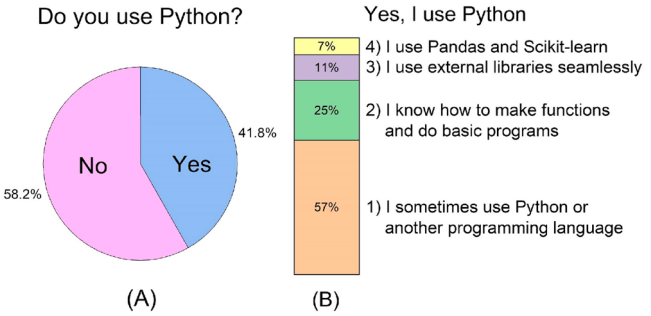


Figure 3. (A) Pie chart indicating the attendance percentage of students exposed previously to Python. (B) Stacked bar chart showing the percentage distribution of previous knowledge concerning Python language.

had never used Python before, and 57% of the remaining 41.8% had very basic knowledge on Python and did not know how to make functions, basic programs, or use external libraries. Prior to the beginning of the workshop, the students were introduced to Google Colab notebooks by means of two videos. In this way, no software installation was required on the students' devices to run the code.

In the first session, a general presentation of the workshop outlined the content, introduced basic machine learning concepts, and mentioned examples of the use of applied machine learning in chemical research and lab work. After this, an introduction to Python was given, together with the reasons for its choice in the workshop. All of the external libraries that would later be used were also introduced, as well as the Google Colab platform and its benefits. After these introductions, a brief 30 min lecture of the concepts of notebook 1 was given, followed by a showcase of the notebook itself and an explanation of the code and live execution of each cell, for approximately 25 min. A similar timeline was followed for the contents of all the subsequent notebooks, including a 5 min break after every notebook to avoid mental strain on the students. Notebooks 1 and 2, and the first topic of notebook 3 were presented on the first day, and the remaining topics of notebook 3 as well as notebooks 4 and 5 were presented on the second day. During and after the sessions, students were encouraged to ask questions via the Google Meet chat and the Discord server. Links to the notebooks and corresponding assignments were made available to the students before each session, and the key to the assignments was sent a week later. After the first implementation of this workshop, the authors noted that interactive and visual execution of code was helpful for students in the introduction to Python programming, that is, using Python Tutor a freely online tool.[34]

https://doi.org/10.1021/acs.jchemed.1c00142
J. Chem. Educ. XXXX, XXX, XXX−XXX

LEARNING OUTCOME

Learning objectives 1 and 2 were fulfilled as students resolved the assignments, probing that they could write and understand simple Python code, and apply machine learning algorithms. A survey helped analyze the effect of this workshop on students' perception toward programming and machine learning. Participation was voluntary and anonymous to minimize biases. Students rated the following four statements on a Likert scale from 5 (Strongly Agree) to 1 (Strongly Disagree):

- Statement A: I think the workshop is useful.
- Statement B: The workshop raised my interest in machine learning.
- Statement C: The workshop raised my interest in programming.
- Statement D: I think the workshop gave me the confidence to continue learning about machine learning and programming independently.

The results of the survey are shown in Figure 4. Different color sections in the plot correspond to the percentage for each answer from A to D, and the number indicates the average from the Likert score for each statement. Students did not answer negatively to the statements and the survey suggests that they attained the expected learning outcomes. The aim of stimulating advanced data analysis is to provide students with the opportunity to learn analytical insights, relevant tools, and programming skills at the undergraduate level. In general, chemistry students, mostly due to a lack of interest or negative feelings, are not keen on computational skills. Therefore, this experience could encourage them to learn new skills and to apply the acquired knowledge in the field.

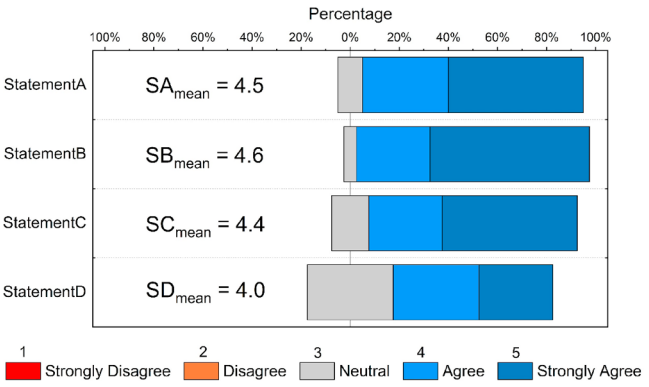


Figure 4. Workshop performance survey using Likert Scale (n = 20).

SUMMARY

A workshop on machine learning with examples of chemical relevance, which was based on five Python notebooks has been presented to complement the undergraduate chemistry curriculum, covering aspects ranging from data modeling for classification and regression to data visualization. The notebooks have been built so that students without programming experience acquire computational skills along the workshop. The hands-on work deals with a real data set of physicochemical characterization of wine, leading to a nourishing experience. Students' response to the workshop was overwhelmingly positive, and their interest in machine learning and programming was awakened.

AUTHOR INFORMATION

Corresponding Author

Diego Onna − *Facultad de Ciencias Exactas y Naturales, Departamento de Química Inorgánica, Analítica y Química Física, Universidad de Buenos Aires, Buenos Aires C1428EGA, Argentina; Instituto de Química Física de los Materiales, Medio Ambiente y Energía (INQUIMAE), CONICET-Universidad de Buenos Aires, Buenos Aires C1428EGA, Argentina;* orcid.org/0000-0002-3158-1915; Email: diego.onna@qi.fcen.uba.ar

Authors

Deborah Lafuente − *Facultad de Ciencias Exactas y Naturales, Departamento de Química Inorgánica, Analítica y Química Física, Universidad de Buenos Aires, Buenos Aires C1428EGA, Argentina;* orcid.org/0000-0003-3660-7894

Brenda Cohen − *Facultad de Ciencias Exactas y Naturales, Departamento de Química Inorgánica, Analítica y Química Física, Universidad de Buenos Aires, Buenos Aires C1428EGA, Argentina;* orcid.org/0000-0002-6550-9014

Guillermo Fiorini − *Facultad de Ciencias Exactas y Naturales, Departamento de Química Inorgánica, Analítica y Química Física, Universidad de Buenos Aires, Buenos Aires C1428EGA, Argentina;* orcid.org/0000-0003-0992-9812

Agustín Alejo García − *Facultad de Ciencias Exactas y Naturales, Departamento de Química Inorgánica, Analítica y Química Física, Universidad de Buenos Aires, Buenos Aires C1428EGA, Argentina;* orcid.org/0000-0002-6536-4935

Mauro Bringas − *Facultad de Ciencias Exactas y Naturales, Departamento de Química Inorgánica, Analítica y Química Física, Universidad de Buenos Aires, Buenos Aires C1428EGA, Argentina; Instituto de Química Física de los Materiales, Medio Ambiente y Energía (INQUIMAE), CONICET-Universidad de Buenos Aires, Buenos Aires C1428EGA, Argentina;* orcid.org/0000-0002-2040-3689

Ezequiel Morzan − *Comisión Nacional de Energía Atómica, Buenos Aires B1950KNA, Argentina*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jchemed.1c00142

Notes

The authors declare no competing financial interest.

Notebooks, Assignments, and solved Assignments in English and Spanish are available at https://github.com/ML4chemArg/Intro-to-Machine-Learning-in-Chemistry (accessed July 2021).

ACKNOWLEDGMENTS

This work has been carried out thanks to the open-source community. We thank students for the feedback provided. The authors thank Leila Morzan and Fiona Britto for the feedback on the graphical abstract. The authors thank Silvana Martin and Anabella Sauer for the proofreading. DO acknowledges the financial support provided by UBACYT20020170200298BA.

REFERENCES

(1) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559* (7715), 547−555.

(2) Gromski, P. S.; Henson, A. B.; Granda, J. M.; Cronin, L. How to explore chemical space using algorithms and automation. *Nature Reviews Chemistry* **2019**, *3* (2), 119−128.

(3) Perkel, J. M. The Internet of Things comes to the lab. *Nature* **2017**, *542* (7639), 125.
(4) Holme, T. A. Can Today's Chemistry Curriculum Actually Produce Tomorrow's Adaptable Chemist? *J. Chem. Educ.* **2019**, *96* (4), 611−612.
(5) Agrawal, A.; Choudhary, A. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *APL Mater.* **2016**, *4* (5), 053208.
(6) Bishop, C. M. *Pattern recognition and machine learning*; Springer: New York, 2006.
(7) Alpaydin, E.. *Introduction to machine learning*; MIT Press: Cambridge, MA, 2020.
(8) Weiss, C. J. Introduction to stochastic simulations for chemical and physical processes: Principles and applications. *J. Chem. Educ.* **2017**, *94* (12), 1904−1910.
(9) Calcabrini, M.; Onna, D. Exploring the gel state: optical determination of gelation times and transport properties of gels with an inexpensive 3D-printed spectrophotometer. *J. Chem. Educ.* **2019**, *96* (1), 116−123.
(10) Kurniawan, O.; Koh, L. L. A.; Cheng, J. Z. M.; Pee, M. Helping Students Connect Interdisciplinary Concepts and Skills in Physical Chemistry and Introductory Computing: Solving Schrödinger's Equation for the Hydrogen Atom. *J. Chem. Educ.* **2019**, *96* (10), 2202−2207.
(11) Langbeheim, E. Simulating the Effects of Excluded-Volume Interactions in Polymer Solutions. *J. Chem. Educ.* **2020**, *97* (6), 1613−1619.
(12) Noyes, K.; McKay, R. L.; Neumann, M.; Haudek, K. C.; Cooper, M. M. Developing Computer Resources to Automate Analysis of Students' Explanations of London Dispersion Forces. *J. Chem. Educ.* **2020**, *97* (11), 3923−3936.
(13) Vargas, S.; Zamirpour, S.; Menon, S.; Rothman, A.; Häse, F.; Tamayo-Mendoza, T.; Romero, J.; Sim, S.; Menke, T.; Aspuru-Guzik, A. Team-Based Learning for Scientific Computing and Automated Experimentation: Visualization of Colored Reactions. *J. Chem. Educ.* **2020**, *97* (3), 689−694.
(14) Hoover, G. C.; Dicks, A. P.; Seferos, D. S. Upper-Year Materials Chemistry Computational Modeling Module for Organic Display Technologies. *J. Chem. Educ.* **2021**, *98* (3), 805−811.
(15) Engelberger, F.; Galaz-Davison, P.; Bravo, G.; Rivera, M.; Ramírez-Sarmiento, C. A. Developing and Implementing Cloud-Based Tutorials That Combine Bioinformatics Software, Interactive Coding, and Visualization Exercises for Distance Learning on Structural Bioinformatics. *J. Chem. Educ.* **2021**, *98* (5), 1801−1807.
(16) Weiss, C. J. Scientific computing for chemists: An undergraduate course in simulations, data processing, and visualization. *J. Chem. Educ.* **2017**, *94* (5), 592−597.
(17) Weiss, C. J. A Creative Commons Textbook for Teaching Scientific Computing to Chemistry Students with Python and Jupyter Notebooks. *J. Chem. Educ.* **2021**, *98* (2), 489−494.
(18) Menke, E. J. Series of Jupyter Notebooks Using Python for an Analytical Chemistry Course. *J. Chem. Educ.* **2020**, *97* (10), 3899−3903.
(19) Kim, S.; Bucholtz, E. C.; Briney, K.; Cornell, A. P.; Cuadros, J.; Fulfer, K. D.; Gupta, T.; Hepler-Smith, E.; Johnston, D. H.; Lang, A. S. I. D.; Larsen, D.; Li, Y.; McEwen, L. R.; Morsch, L. A.; Muzyka, J. L.; Belford, R. E. Teaching Cheminformatics through a Collaborative Intercollegiate Online Chemistry Course (OLCC). *J. Chem. Educ.* **2021**, *98* (2), 416−425.
(20) Janet, J. P.; Kulik, H. J. *Machine Learning in Chemistry*; American Chemical Society: Washington, DC, 2020.
(21) Joss, L.; Müller, E. A. Machine learning for fluid property correlations: classroom examples with MATLAB. *J. Chem. Educ.* **2019**, *96* (4), 697−703.
(22) Antonelli, T. M.; Olivieri, A. C. Developing and Implementing an R Shiny Application to Introduce Multivariate Calibration to Advanced Undergraduate Students. *J. Chem. Educ.* **2020**, *97* (4), 1176−1180.
(23) The Carpentries. https://carpentries.org (accessed July 2021).
(24) MolSSI Education Resources. http://education.molssi.org/resources.html (accessed July 2021).
(25) Bisong, E. Google colaboratory. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform*; Apress: Berkeley, CA, 2019; pp 59−64.
(26) Iqbal, S.; Bhatti, Z. A. A qualitative exploration of teachers' perspective on smartphone usage in higher education in developing countries. *International Journal of Educational Technology in Higher Education* **2020**, *17* (1), 1−16.
(27) Hossain, M. E.; Ahmed, S. Z. Academic use of smartphones by university students: a developing country perspective. *Electronic Library* **2016**, *34*, 651.
(28) Introduction to Machine Learning for chemists. https://github.com/ML4chemArg/Intro-to-Machine-Learning-in-Chemistry (accessed July 2021).
(29) Cortez, P.; Cerdeira, A.; Almeida, F.; Matos, T.; Reis, J. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems* **2009**, *47* (4), 547−553.
(30) Wine Quality Data Set. https://archive.ics.uci.edu/ml/datasets/wine+quality (accessed July 2021).
(31) Danjou, P. E. Distance teaching of organic chemistry tutorials during the COVID-19 pandemic: Focus on the use of videos and social media. *J. Chem. Educ.* **2020**, *97* (9), 3168−3171.
(32) Hamer, M.; Beraldi, A. M.; Gomez, S. G. J.; Ortega, F.; Onna, D.; Hamer, M. Glowing-in-the-Screen: Teaching Fluorescence with a Homemade Accessible Setup. *J. Chem. Educ.* **2021**, *98* (8), 2625−2631.
(33) Mayer, M.; Baeumner, A. J. A megatrend challenging analytical chemistry: biosensor and chemosensor concepts ready for the internet of things. *Chem. Rev.* **2019**, *119* (13), 7996−8027.
(34) Visualize Code Execution. http://pythontutor.com/ (accessed July 2021).

https://doi.org/10.1021/acs.jchemed.1c00142
J. Chem. Educ. XXXX, XXX, XXX−XXX