

Teaching Basic Concepts in Machine Learning to Engineering Students: A Hands-on Approach

Dr. David Olubiyi Obada, Ahmadu Bello University, Nigeria

David O. Obada holds a Ph.D. degree in mechanical engineering from the Ahmadu Bello University, Zaria, Nigeria, specializing in production/industrial engineering. His research interests include fracture mechanics, advanced materials, and condensed matter physics. Before joining the Atlantic Technological University, Ireland, David was a research fellow at the University of Ghana, National Environmental Engineering Research Institute, Nagpur, India, and the University of Birmingham, UK. Also, David was a research and teaching fellow at the Massachusetts Institute of Technology (MIT), USA, and holds a Kaufmann Teaching Certificate from MIT.

Mr. Simeon Akindele Abolade, Atlantic Technological University, Ireland

Simeon Akindele Abolade is a PhD student at the Atlantic Technological University (ATU) in Sligo, Ireland, and a member of the Mathematical Modelling and Intelligent Systems for Health and Environment (MISHE) research group. He holds a BSc (Hons) degree in Physics from the University of Ilorin, Nigeria, and MSc in theoretical physics from the University of Ibadan, Nigeria. He is currently on his PhD programme at ATU, Sligo, under the supervision of Dr Akinlolu Akande, Prof. Stefano Sanvito and Dr Fedwa El-Melouhi. His research interests include modeling novel materials for photovoltaics, photocatalytic, and thermoelectric applications using state-of-the-art computational tools based on Density Functional Theory and Machine Learning.

Mr. Shittu Babatunde Akinpelu, Atlantic Technological University, Ireland

Shittu Babatunde Akinpelu is a Ph.D. student at Atlantic Technological University (ATU) in Ireland. He completed his bachelor's degree in physics at the Federal University of Technology (FUTA), Akure, Nigeria. While pursuing his master's degree in condensed matter physics in 2021, Babatunde gained valuable research experience as a research assistant at FUTA. In 2022, Babatunde joined the Modelling and Computation for Health and Society (MOCHAS) cohort at ATU, Ireland for his Ph.D. His current research focuses on material modeling using machine learning and soft computing, with a particular interest in discovering novel materials with unique properties that could be applied in the fields of energy and health.

Ayodeji Nathaniel Oyedeji, Ahmadu Bello University, Nigeria

Ayodeji Nathaniel Oyedeji is an advanced manufacturing and materials science researcher. With exceptional performances in both his B.Sc and M.Sc studies in Mechanical Engineering at Ahmadu Bello University, Nigeria, he is now pursuing a PhD in Industrial Engineering at Stellenbosch University, South Africa. Alongside his scholarly pursuits, Ayodeji demonstrates a keen interest in engineering education. He has made significant contributions to his field through a prolific publication record and active participation in academic conferences. Possessing a diverse skill set, including strong communication abilities and analytical proficiency, Ayodeji is also an avid reader and enjoys nature. His trajectory reflects a commitment to continuous growth and making a meaningful impact within engineering and beyond.

Dr. Emmanuel Okafor, King Fahd University of Petroleum and Minerals, Saudi Arabia

Emmanuel Okafor holds a Ph.D. in Artificial Intelligence from the University of Groningen, Netherlands, specializing in computer vision, machine learning, and reinforcement learning. His research interests include medical informatics, robotics, animal monitoring, and prediction of biomaterial properties. Before joining the King Fahd University of Petroleum and Minerals, Saudi Arabia, Emmanuel worked as a faculty member at the Department of Computer Engineering, Ahmadu Bello University, Nigeria. Furthermore, Emmanuel was a research and teaching fellow at the Massachusetts Institute of Technology (MIT), USA, and earned a distinction in the course: "An Introduction to Evidence-Based Undergraduate STEM Teaching" coordinated by the Center for the Integration of Research Teaching and Learning (CIRTL), 2022.

Ms. Cynthia Ujuh Odili, Ahmadu Bello University, Nigeria

Cynthia U. Odili is a Ph.D. student at the Department of Water Resources and Environmental Engineering at Ahmadu Bello University (ABU). She completed her bachelor's degree in Chemical Engineering at the Federal University of Technology, Minna, Nigeria and a master's degree in Water Resources and Environmental Engineering at ABU, Zaria. As a researcher, she is passionate about finding sustainable and climate-smart solutions to global water challenges. She has a keen interest in Engineering Education and she is also a student at the African Centre of Excellence on New Pedagogies in Engineering Education (ACENPEE), ABU, Nigeria. Her research interests include optimization of process parameters, machine learning, bioremediation, biomaterials, environmental pollution, and water treatment.

Vanessa Faustina Ogenyi

Mr. Sokoga Victor Ategbe, Ahmadu Bello University, Nigeria

An undergraduate student at the Department of Chemical Engineering at Ahmadu Bello University (ABU) in Zaria, Nigeria.

Prof. Adrian Oshioname Eberemu, Ahmadu Bello University, Nigeria

Adrian Oshioname EBEREMU is a Professor of Geotechnical/Geoenvironmental Engineering in the Department of Civil Engineering Ahmadu Bello University Zaria, Nigeria. He has been an academic for more than 18 years, before which, he spent his earlier years practicing as a geotechnical and civil engineer in the industry in Nigeria. He is a product of the DSC Technical High School, the 90set where a good academic foundation was laid. He had his Bachelor of Engineering in Civil Engineering from Federal University of Technology Owerri, FUTO in 1997; MSc and PhD in 2003 and 2008, respectively from the Ahmadu Bello University Zaria.

He has taught and examined students both at the undergraduate and post graduate levels in several Nigerian Universities such as University of Maiduguri, Modibbo Adama University of Technology Yola, Kano State University of Technology, University of Agriculture Makurdi, Bayero University Kano, Abubakar Tafawa Balewa University Bauchi, Nnamdi Azikwe University Awka and Covenant University Ota, Federal University of Technology Minna, Michael Okpara University of Agriculture Umudike, Olusegun Agagu University of Science Technology Okitipupa and Universiti Teknologi Malaysia.

His primary area of expertise are in geo-material site characterization, deep foundation, the beneficial reuse of waste materials in soil improvement, solutions to geo-environmental problems, waste containment barriers and covers, Biogeochemical Processes in Geotechnical Engineering (Microbial Induced Calcite Precipitation) and unsaturated soils (collapsible soils) and lately engineering education. He has many published works in peer-reviewed journals, conference proceedings and chapters as well as technical reports to his credit in the various research area.

He is currently the academic and research coordinator with the African Center of Excellence on New Pedagogies in Engineering Education (ACENPEE), Ahmadu Bello University Zaria; a World Bank funded Development Impact project with the aim of scaling up post graduate education at the MSc/PhD levels through regional specialization and collaboration in the West African Sub-region.

Adrian is a registered Engineer with Council for Regulation of Engineering Practice in Nigeria (COREN), a member of the Nigerian Society of Engineers, a member of the American Society for Civil Engineers as well as the International Society for Soil Mechanics and Geotechnical Engineering (ISSMG).

Fatai Olukayode Anafi, Ahmadu Bello University, Nigeria

Professor of Mechanical Engineering at Ahmadu Bello University, Zaria Nigeria. His research interests are in Energy, Thermo-fluids, Engineering Education, Project and Operations Management. He has over 50 refereed publications to his credit, attended National and International Conferences and has 31 years teaching, research and administrative experience. He is the M & E officer of Africa Centre of Excellence on New Pedagogies in Engineering Education (ACENPEE), a World Bank ACE Impact Project. He is a recipient of RMRDC Nigeria Research Grants, World Bank Science and Technology Education Post-Basic (STEP-B) and Africa Centre of Excellence Research Grants, among others.

Abdulkarim Salawu Ahmed, Ahmadu Bello University, Nigeria
Dr. Akinlolu Akande, Atlantic Technological University, Ireland

Dr. Akinlolu Akande is a computational material scientist and lecturer at the Atlantic Technological University (ATU) Sligo, Ireland. He is a Principal Investigator in the Mathematical Modelling and Intelligent Systems for Health and Environment (MISHE) and Modelling Computation for Health And Society (MOCHAS) Research Groups at the ATU. He completed his PhD in Physics at Trinity College Dublin, Ireland, and subsequently worked as a postdoctoral research fellow at the same institution. During this time, he combined research in computational material sciences with teaching duties in undergraduate laboratories. He then served as an assistant lecturer at the Dundalk Institute of Technology in Dundalk, Ireland, before joining the Institute of Technology Sligo (now ATU Sligo). Akinlolu is a Senior Fellow of the Higher Education Academy (SFHEA), a recognition of his expertise in teaching and learning in higher education.

Teaching basic concepts in machine learning to engineering students: A hands-on approach

David O. Obada^{1,2,10,11*}, Simeon A. Abolade², Shittu B. Akinpelu³, Ayodeji N. Oyedeji¹, Emmanuel Okafor⁴, Cynthia U. Odili⁵, Vanessa F. Ogenyi⁶, Victor S. Ategbé⁸, Adrian O. Eberemu^{7,10}, Fatai O. Anafi^{1,10}, Abdulkarim S. Ahmed^{8,10}, Raymond B. Bako^{9,10}, Akinlolu Akande^{2,3}

¹Department of Mechanical Engineering, Ahmadu Bello University, Zaria, 810222, Nigeria.

²Mathematical Modelling and Intelligent Systems for Health and Environment Research Group, School of Science, Atlantic Technological University, Sligo, F91 YW50, Ireland.

³Modelling and Computation for Health and Society, Atlantic Technological University, Sligo, F91 YW50, Ireland.

⁴SDAIA-KFUPM Joint Research Center for Artificial Intelligence, King Fahd University of Petroleum and Minerals, Dhahran, 31261, Saudi Arabia.

⁵Department of Water Resources and Environmental Engineering, Ahmadu Bello University, Zaria, 810222, Nigeria.

⁶Department of Polymer and Textile Engineering, Ahmadu Bello University, Zaria, 810222, Nigeria.

⁷Department of Civil Engineering, Ahmadu Bello University, Zaria, 810222, Nigeria.

⁸Department of Chemical Engineering, Ahmadu Bello University, Zaria, 810222, Nigeria.

⁹Department of Educational Foundations and Curriculum, Ahmadu Bello University, Zaria, 810222, Nigeria.

¹⁰Africa Centre of Excellence on New Pedagogies in Engineering Education, Ahmadu Bello University, Zaria, 810222, Nigeria.

¹¹Multifunctional Materials Laboratory, Shell Office Complex, Department of Mechanical Engineering, Ahmadu Bello University, Zaria, 810222, Nigeria.

Corresponding author: David O. Obada (doobada@abu.edu.ng)

Abstract

According to a recent survey conducted by the Corporate Member Council of the American Society of Engineering Education (ASEE), there exists a notable disparity in the skill sets of engineering graduates about Artificial Intelligence (AI). To address this disparity from the African context, the Africa Centre of Excellence on New Pedagogies in Engineering Education organized a machine learning (ML) workshop for engineering students from different disciplines. Seventy-three (73) students enrolled for the workshop and the modules covered during this workshop were: Introduction to ML Models, ML Frameworks, Additive Explanations in ML, Performance Metrics, and Introduction to Ensemble Learning Techniques. The hands-on session involved the use of categorical boosting (CatBoost), an ensemble learning technique, to predict the bulk modulus, a mechanical property, of 199 ABX₃ perovskite materials which was used as a problem set. The input features influencing the CatBoost model decisions were subsequently established. Correlation analysis on the input feature space removed features with high collinearity. The SHapley Additive exPlanations (SHAP) was used to analyze the decision-making rationale of the model. Evaluation of the model performance based on the coefficient of determination R² value (0.94) revealed that the model demonstrates good performance in predicting the bulk modulus of the perovskite materials used during the practical sections. The survey results after the teaching and practical sessions indicate that the

learning modules are an effective introduction for novice engineering students in this domain and raise awareness of the importance of this important sub-section of AI.

Keywords: Engineering Education; Artificial Intelligence; Machine Learning; Perovskites; Materials Science

1. Introduction

Machine learning (ML) is a subfield of artificial intelligence (AI) that has been effectively applied in various problem domains such as computer vision [1], speech recognition [2], machine translation [3], fault detection [4], predictive maintenance [5], robotic tactile sensing [6], and social media analysis [7]. It has become relevant in several sectors, allowing individuals with diverse educational backgrounds to utilize it, rather than being limited to computer science or mathematics courses [8]. Therefore, it is vital to familiarize young individuals with ML at an early stage to enable their social and economic engagement, particularly considering the increasing need for ML experts [9]. ML is primarily taught at higher education institutions, particularly at the university level. Typically, it is exclusively included in specialized computer science or data science programs [10,11] and is a standard component of computing courses in higher education as recommended by curriculum guidelines [12]. The instruction of ML ideas encompasses the primary stages of a process centered around human involvement in the development of its applications, including data preparation, model training, and performance evaluation [13]. At present, there is a growing number of instructional materials to teach basic ML concepts to high-level students targeting mainly beginners [14,15]. Nevertheless, there appear to be inadequate courses that offer a wider range of interdisciplinary integrations, as well as a scarcity of support information for students who possess a strong interest in the fundamental principles.

Given the increasing popularity of ML, there is a growing demand for educational offerings in this area. To bridge this gap, the Africa Centre of Excellence on New Pedagogies in Engineering Education (ACENPEE) has initiated several training workshops in AI. ACENPEE is one of Nigeria's 17 World Bank-supported projects for Development Impact (ACE Impact Project). The Africa Higher Education Centers of Excellence (ACE) Project is an initiative by the World Bank in collaboration with participating countries' governments to support higher education institutions specializing in Science, Technology, Engineering, and Mathematics (STEM), Environment, Agriculture, applied Social Science/Education, and Health. ACENPEE, among other things, is responsible for developing and providing a world-class teaching and learning environment that stimulates and promotes innovation in technopedagogical skills and competencies for engineering education and practices. ACENPEE's mandate is to fill the gap that exists in the training of engineering professionals where there is over-reliance on traditional teaching methods. The key educational and applied research goal of ACENPEE is to use new pedagogies to enhance the training of engineering professionals with the capacity to identify existing challenges and provide solutions through high-level research.

Therefore, a motivation for ACENPEE for the first series of workshops in ML is centered around the fact that the availability of ML courses for engineering students in developing countries may be limited, as most options are offered through computer science departments. However, these courses primarily focus on theory rather than practical applications. Additionally, they are often overcrowded due to the high number of students enrolling for the course. Also, the available education research literature on teaching ML for non-computer

science majors is quite limited [16]. Most of the current education literature and teaching materials are primarily tailored for computer science students [17]. Sulmont et al. [16] in their study interviewed instructors who teach ML to non-majors. The instructors expressed a common belief that ML is difficult for individuals who lack expertise in statistics and programming. This belief was also shared by the participants in their study. Reyes et al. [18] developed a graphical tool that familiarizes advanced students with ML, even if they lack extensive programming or mathematical expertise. The authors achieved this by proposing a visualization tool for high school students to introduce them to ML using gamification concepts and curriculum adaptation. In addition, Huang et al. [19] developed active learning labware that allows first-year undergraduate engineering students to engage with real-world problem-solving using ML. Their pre-post survey evaluation of the teaching/learning tool reveals user experiences and effectiveness, guiding further development. Given the importance of pedagogical aspirations, it is essential to recognize the significant obstacles that may hinder the successful implementation of ML modules for novice engineers. An example is the use of ML libraries to conduct ML tasks. For instance, the Scikit-learn framework in Python [20] offers a diverse selection of ML models that are enclosed behind a consistent programming interface. However, learners need to have a fundamental understanding of ML libraries and codes before they can engage in experimentation and instructors can develop their curricula along these lines to prepare the future generation of scientists with strong analytical abilities [21–23]. These were taken into consideration for building the basic ML course contents (Appendix A) used for the participants during the training facilitated by some of the authors. Furthermore, it is worth noting that when developing an ML curriculum, it is often desirable to incorporate programming exercises that allow students to gain practical experience in solving real-world problems. The Python code has become a popular choice for introductory-level programming courses in the physical sciences due to its widespread use in scientific research [24–26] with increasing trends of using digital notebooks to write and execute this code. These notebooks serve as multimedia tools that enhance code readability and reproducibility. They also provide a user-friendly introduction for students. These files are commonly given to users as standalone files (.ipynb extension), allowing them to be executed either locally or on the cloud. Services like Google Colaboratory (Google Colab) [27] have been noted for their ability to offer a consistent and fair user experience [23]. The authors have utilized this platform to study the electronic band gaps of ABX_3 perovskite materials [28] and this dataset extension was utilized to construct ipynb files (snippet displayed in Appendix C) to practically illustrate the module course content.

Despite several innovative advances in teaching ML to non-majors globally, there is a need to further emphasize teaching ML to non-majors from an African context to match contemporary skill sets. To address this issue of limited access to ML for non-majors, a workshop was arranged by ACENPEE with the specific aim of instructing and inspiring the participants who were mostly engineering students in the application of the basic concepts of ML. The process of materials discovery, which involves the identification of new materials with specific features, served as the foundation for instructing the participants on the fundamental concepts of ML [29].

Therefore, this paper serves to report the module design and a hands-on technique that was successfully implemented by ACENPEE to help students of various engineering backgrounds develop self-efficacy in ML. The next sections describe the approach used for the workshop, the discussion of students' perceptions of the learning experience assessed through the learner's

satisfaction survey, as well as the concluding section. The designed modules and snippets of the scripts used during the workshop are described in the appendix section.

2. Platforms, tools, and hands-on outcomes

The workshop on basic concepts in ML for materials discovery was held on the 6th-13th of October 2023, at ACENPEE, Ahmadu Bello University, Zaria, Nigeria. The workshop enrolled a total number of 73 students both online and onsite (hybrid). The participants were drawn from several universities in Nigeria and the African region. The workshop, among other things, aimed to bridge the gap between traditional materials discovery and the transformative potential of ML. The workshop targeted undergraduate and postgraduate students as well as researchers with a passion for innovation looking to develop skills related to ML and to improve their research competence. For this program, an introduction to ML frameworks, feature engineering, SHapley Additive exPlanations (SHAP), performance metrics used for ML model evaluation, and an introduction to ensemble learning techniques were explored (Appendix A).

Participants also explored the essential tasks of ML from the supervised learning models to unsupervised and reinforcement learning. They were also exposed to using ML notebooks on *Google Colab*, which incorporates various programming languages such as Python, Scipy, C++, R, amongst many others, bridging the gap between theory and practice.

The key objectives of the ML workshop were:

1. to provide a collaborative platform for knowledge exchange fostering interdisciplinary action.
2. to teach the participants how to use popular ML libraries such as TensorFlow or Scikit-learn, and codes such as Python.
3. to help the participants apply ML to their research work and solve real-world problems.

The facilitators provided a guide on the core principles of ML algorithms and hands-on practicals with datasets on the mechanical properties of perovskite materials which was used in developing the problem sets for the learners. The mechanical properties (bulk modulus) prediction of 199 examples of ABX₃ perovskite compounds were used to demonstrate to the learners how to practically apply the contents taught during the teaching session. The datasets were split into train and test examples as usual for most ML tasks. For context, the bulk modulus (the target output) was selected because it is a physical parameter that measures a material's resistance to bulk compression. The datasets used during this session are extensions of those reported by Obada et al. [28] and Korbel et al. [30] for predicting the bandgaps of 199 cubic perovskite compounds. 17 input features (Pauling electronegativity, Covalent radius, first ionization energy, Elastic constants, volume per atom, indirect band gap, row of elements, and the atomic radius of elements) related to the prediction of the bulk modulus were carefully selected. Feature engineering was used to remove the redundant features reducing the features to 12 (as depicted in Figure 2b) with the aid of the Pearson Correlation Coefficient (PCC) [31,32] as shown in Figure 1. This demonstrated the importance of feature selection for improving the predictive power of ML models to the participants.

To discuss some of the outcomes of the hands-on session, the performance metrics for the training and test data sets, obtained through fivefold cross-validation, exhibited clear patterns of predictive performance. The testing phase correlation plot is shown in Figure 2a. The coefficient of determination (R^2) obtained from the performance metric values indicates a strong model fit for both the training data (0.99) and the test data (0.94). The R^2 values of the

models used in the hands-on session demonstrates the effectiveness of the feature engineering process in capturing a significant amount of variance in the data. It is safe to say that the practical session enhanced the learner's comprehension of the predictive abilities of ML models. The SHAP analysis (Figure 2b) offered an explanatory framework for understanding the importance of the input features for different ML methods [28]. This approach was employed to explain the influence of each input feature used in predicting the bulk modulus of the perovskite compounds. The top-ranked feature is situated at the apex, while the hierarchy of features descends along the feature axis by their respective significance as shown in the SHAP plots. The facilitators explained how each feature contributes significantly to the prediction of bulk modulus. From the choice of features to the student's interpretation from a physics and materials science standpoint, the SHAP plots explain the bulk modulus which is a physical property that assesses a compound's ability to withstand bulk compression. It can be described as the quantitative measure of the ratio of applied pressure to the resulting strain on a specific material. In general, smaller atomic and covalent radii result in stronger covalent bonds and larger bulk modulus. This is because smaller atoms can often develop closer bonding distances and greater bond formation, resulting in resistance to compression. This is supported by Figure 2b, which shows the SHAP plots for the bulk modulus. All the features shown in Figure 2b are important to predict the bulk modulus of perovskite compounds, however, some of the features are more important (situated at the apex). This means that Pauling electronegativity, covalent and atomic radii, and rows of the elements in the periodic table have a high impact on bulk modulus prediction. The most important feature which is the Pauling electronegativity (Figure 2b) is based on dissociation energies and is not a feature of individual atoms, but rather of linked atoms. As a result, the energy of these perovskite compounds is often derived from electron interactions within the elemental composition, which can reinforce and improve the materials' mechanical characteristics. For the features at the base of the SHAP plots, for instance the bandgap, this feature affects the bulk modulus because this property is determined by interatomic bonding and atomic arrangement inside the material, both of which are impacted by the electronic crystal structure.

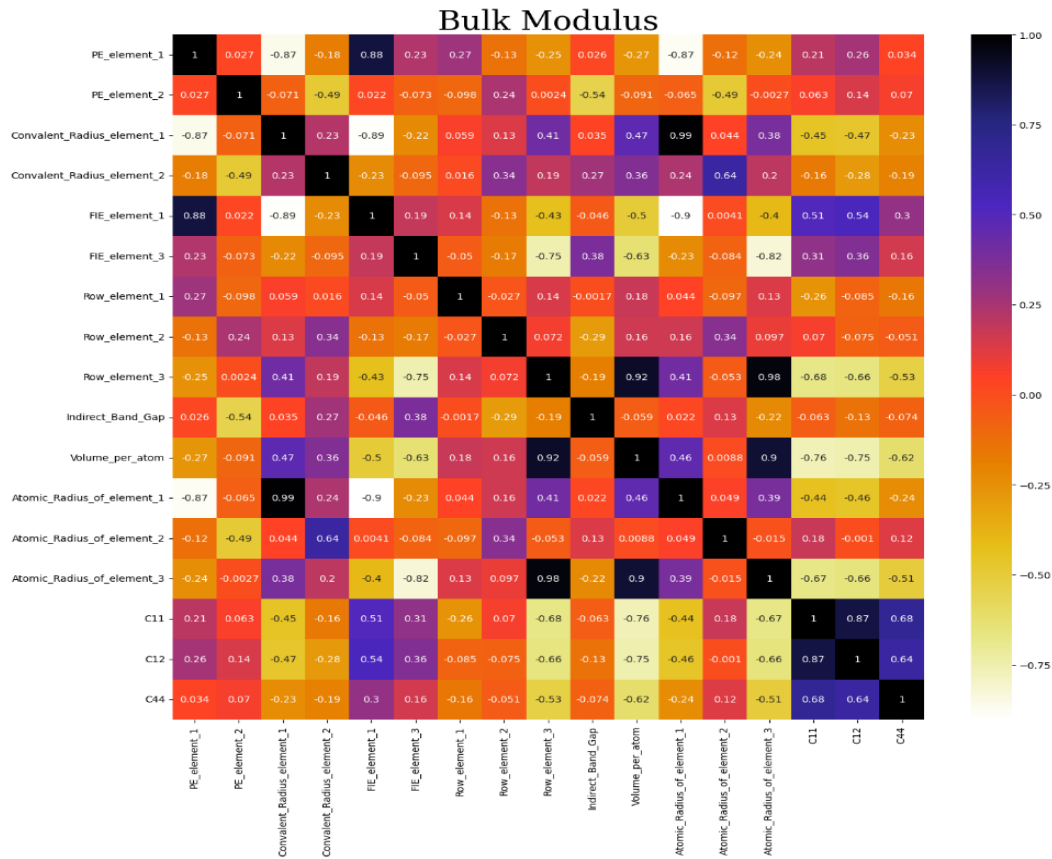


Figure 1: Input features correlation plot heatmap for predicting the bulk modulus of ABX_3 perovskites: C_{11} , C_{12} , and C_{44} (Elastic constants); PE: Pauling electronegativity; FIE: First ionization potential.

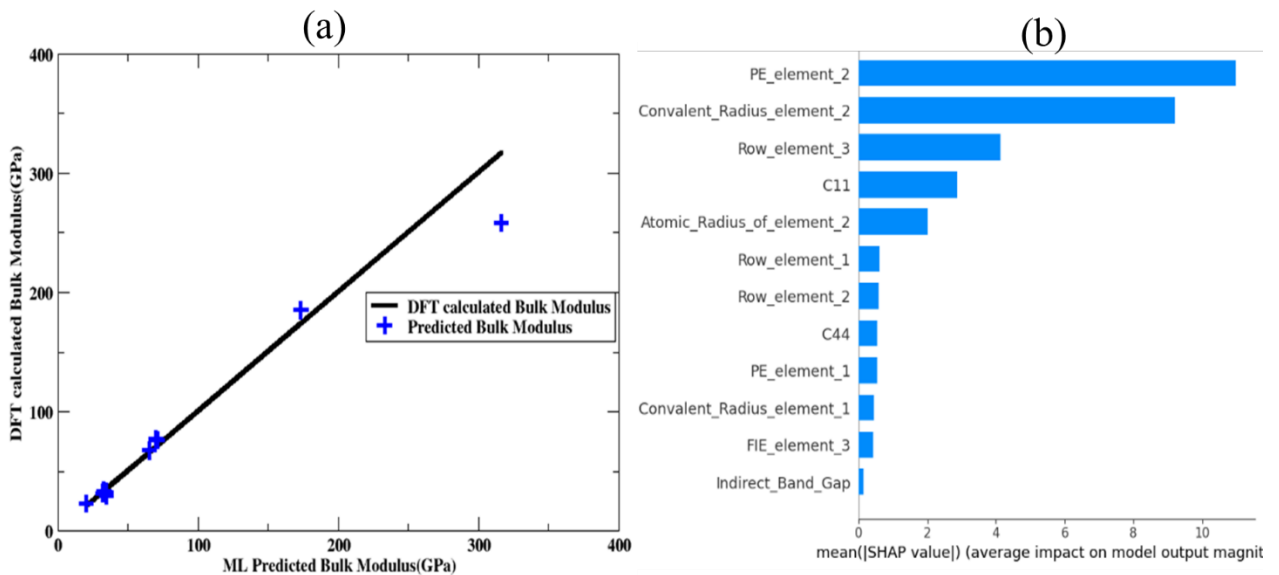


Figure 2: (a) The correlation plots for the testing phase of the problem set used during the workshop, and (b) SHAP analysis for the target output (bulk modulus) of the perovskite compounds.

At the end of the program, a learner’s satisfaction survey was carried out. A total of 49 students completed the survey. All students who enrolled in the workshop were emailed an invitation to participate in this online survey using Google Forms, and the respondent’s results were taken and recorded. Respondents were asked to rate the information obtained from the training, their use of knowledge obtained from the training, the facilitator’s knowledge of the topics covered, the speaker’s presentation skills, the content of slides and virtual aids, future enrollment for training, sessions expectation, overall training evaluation, and recommendation of training sessions to colleagues. In the section that follows, the survey reports are described.

3. Learner’s satisfactory survey

A survey was administered to all participants. Out of 73 participants who registered for the workshop, 49 respondents participated in the survey. Participants were given the freedom to choose the percentage of the information they obtained from the training, the use of knowledge obtained from the training, the probability of them registering for this training in the future, the speaker’s knowledge of the topics covered, the speaker’s presentation skills, the content of slides and virtual aids, session’s expectations, training evaluations, and recommendation of the training sessions to their colleagues. The survey’s findings are displayed in Figures 3-11.

3.1.1 Percentage of information obtained from the ML workshop.

Figure 3 depicts the percentage of information obtained by the participants from the training sessions. The survey conducted after the training shows that 55% of the participants were able to comprehend at least 75% of the information presented during the workshop. However, none of the respondents reported obtaining no information from the training sessions. This indicates that all the respondents gained helpful information from the training. This indicates that the training provided in the workshop was well-received by the participants.

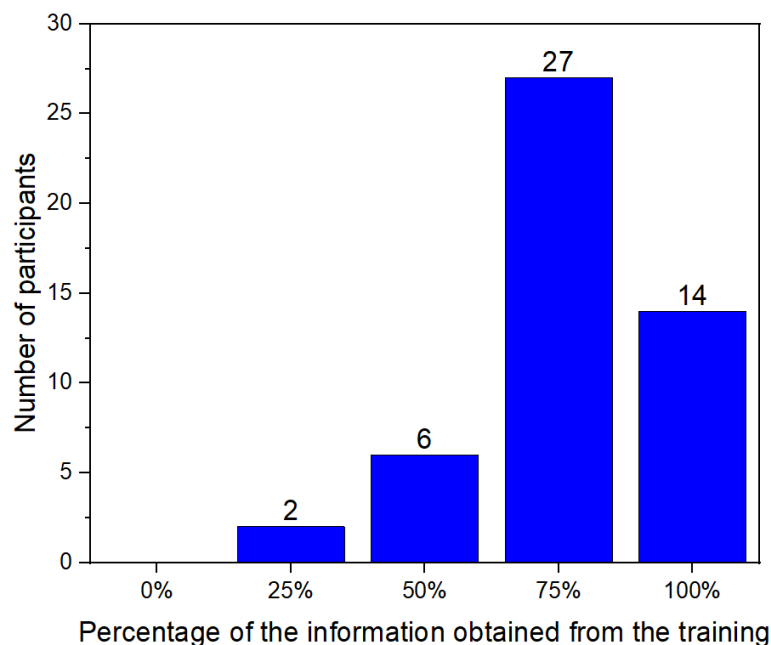


Figure 3: Respondents' feedback on the information obtained from the training.

3.1.2 The use of knowledge obtained from the training.

The tendency of the participants to use the knowledge obtained from the training was also evaluated using the learner satisfaction survey. The response of the respondents is depicted in Figure 4. From their responses, an average of 41% agreed to use the knowledge obtained from the ML workshop immediately, while 31% agreed to employ this knowledge in 2 to 6 months. In addition, 10% agreed to utilize it in 7 to 12 months, and 18% agreed to use this information in the future. This signifies that most of the participants found the training useful and have identified potential areas where the knowledge obtained from the workshop will be useful.

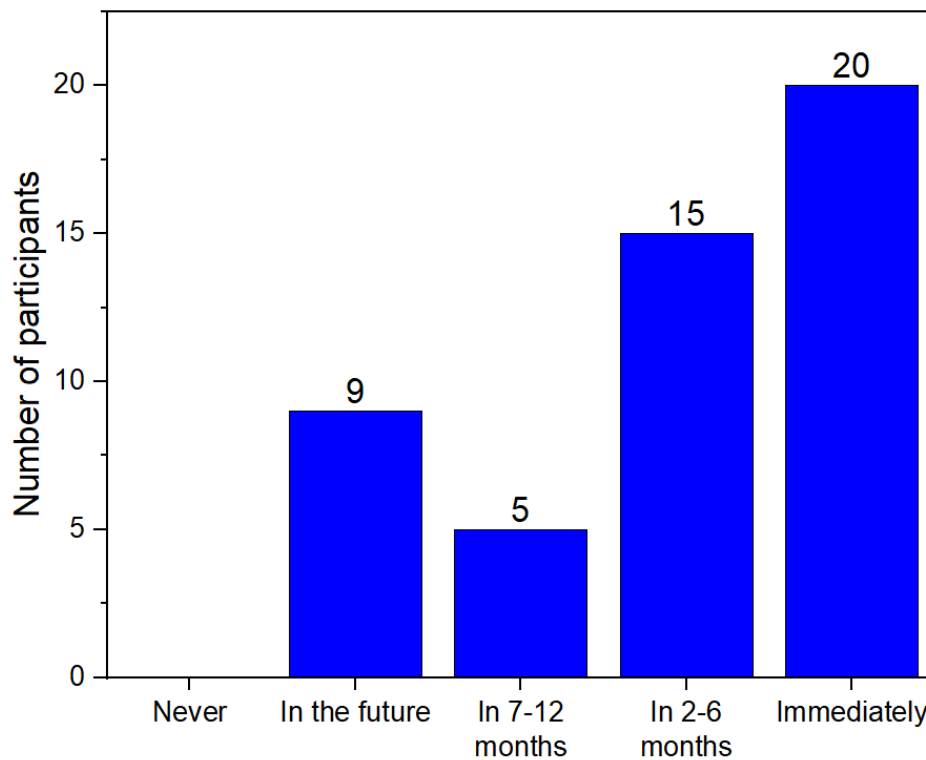


Figure 4: Respondents' feedback on the use of knowledge obtained from the training.

3.1.3 Registering for such training in the future.

Figure 5 shows the probability of respondents registering for such training in the future. Out of 49 respondents, 98% indicated a Yes to registering for this type of training, while 2% indicated a Maybe to registering for this type of training. Also, it was observed that none of the respondents chose "No" revealing a high level of learners' interest in the workshop. One possible reason for the responses could be that the participants regarded the impact of the training as worthwhile and one which has significantly boosted their interest in ML and this has encouraged further studies in this area of scientific importance.

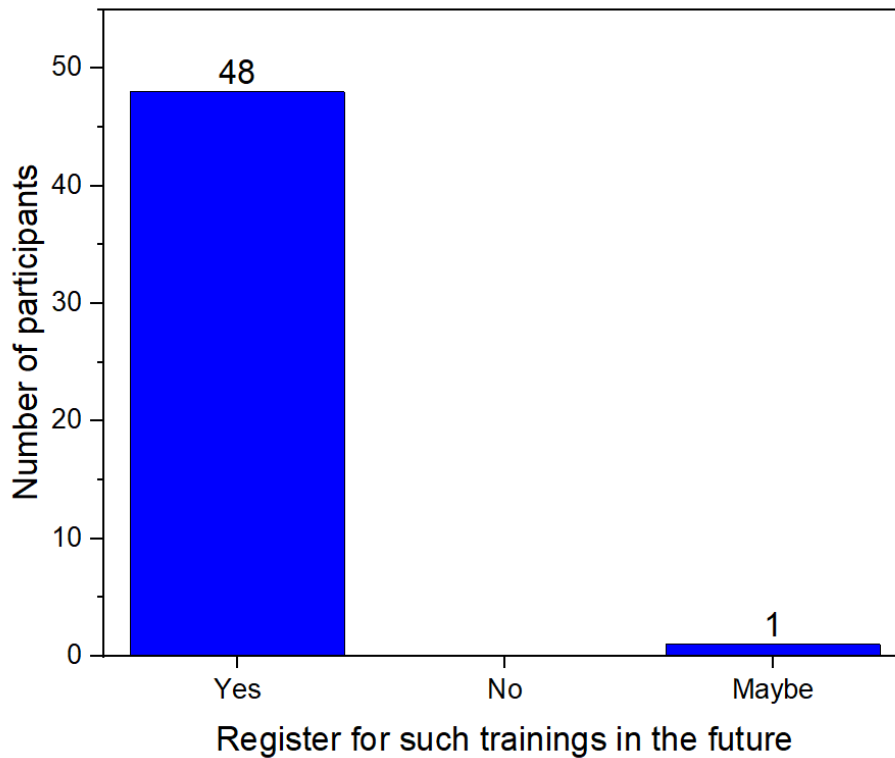


Figure 5: Respondents' feedback on possible registration for future training

3.1.4 The speakers' knowledge of the topics covered.

The participants were exposed to various topics on the basic concepts in ML for materials discovery. The facilitator's knowledge of the topics covered was rated by the respondents in this survey as shown in Figure 6. From the results obtained, 78% of the respondents rated the facilitators' knowledge as excellent, while 22% rated their knowledge as good. This suggests that during the training, the hands-on activities, the facilitators' didactics, attentive interaction, and the visual tools used, impressed the participants and improved their understanding of the subject area. Another implication from the feedback could be that the interaction modes effectively addressed learners' questions making them suggest that the facilitators' knowledge of the subject area was excellent.

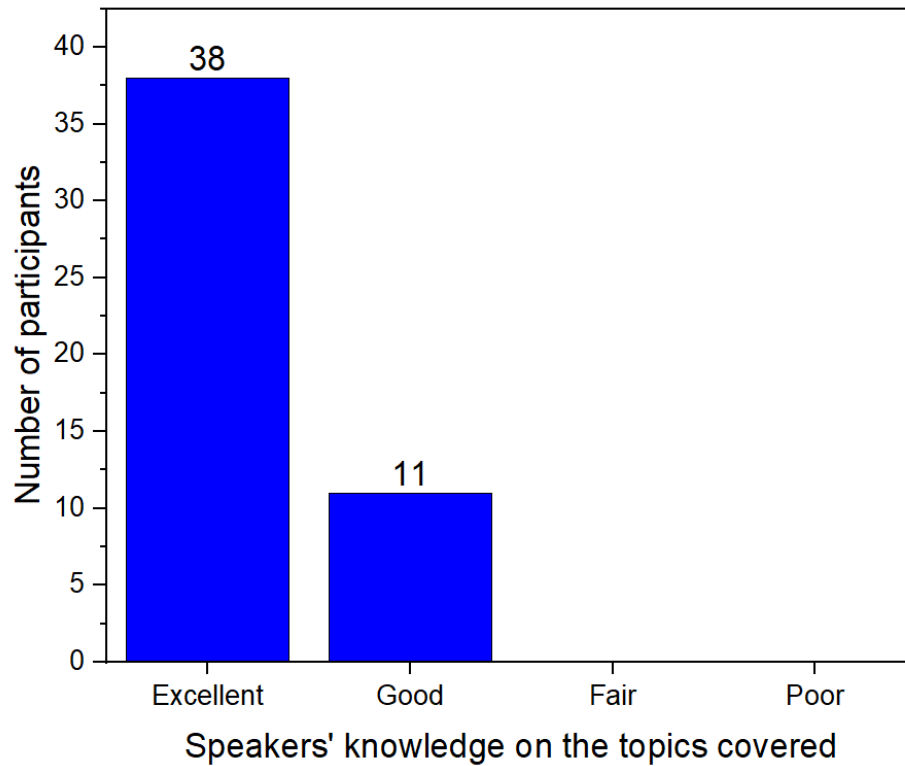


Figure 6: Respondents' feedback on the speaker's knowledge of the topics covered.

3.1.5 Speaker's presentation skills

The results obtained from the survey indicated that over 61% of the respondents rated the speakers' presentation skills as excellent, 37% as good, and 2% as fair as shown in Figure 7. This suggests that most of the participants expressed the speakers' presentation skills as very fascinating and easy to understand.

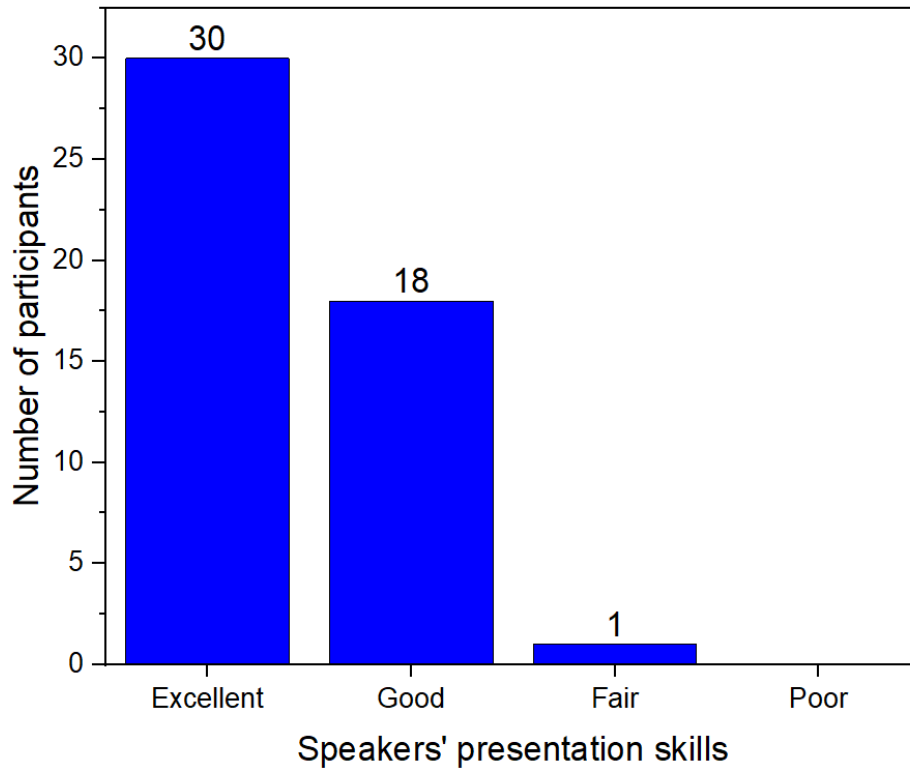


Figure 7: Respondents' feedback on the speaker's presentation skills

3.1.6 Content of slides and virtual aids

The content of slides and virtual aids used by the facilitators was assessed by the participants and their responses are shown in Figure 8. From the results obtained, 47% of the respondents rated the contents of the slides (snippets shown in Appendix B) and virtual aids as excellent, while 51% rated the content of slides and virtual aids as good. However, 2% of respondents rated the materials as fair suggesting that participants found the contents of the slides and materials used for carrying out the teaching as clear and systematic. The training aimed at beginners could benefit from additional optional materials to provide in-depth information on certain topics. This could also motivate further training, such as introducing other Python-based ML models or guiding students in more problem sets on ML.

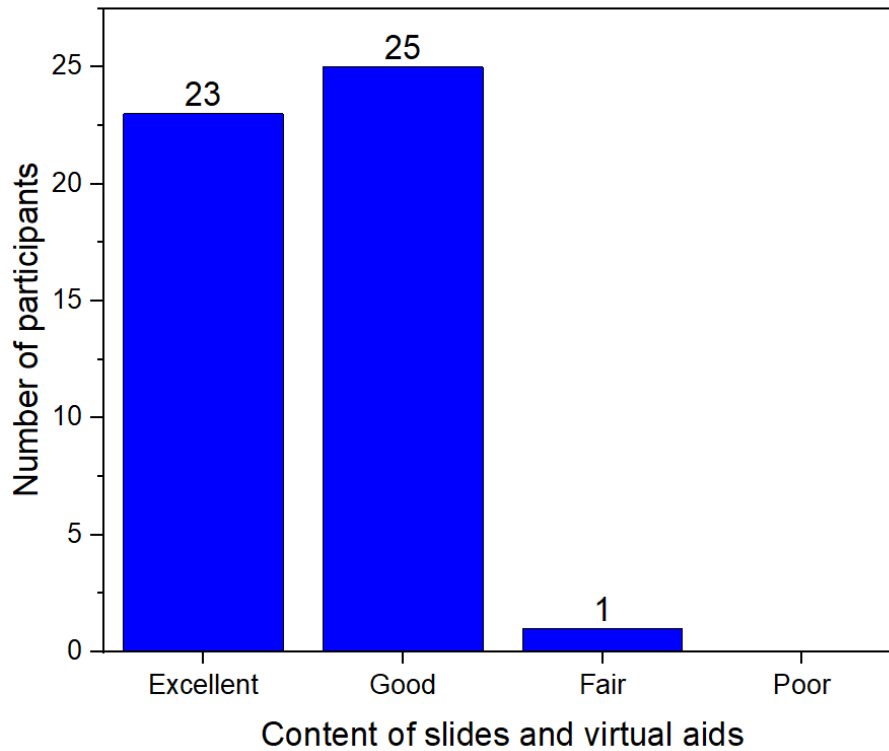


Figure 8: Respondents' feedback on the contents of the slides and virtual aids used during the training sessions.

3.1.7 Session expectations

Participants' knowledge of how the workshop met their expectations was also assessed and their responses on their understanding of each of the sessions are shown in Figure 9. The results from the survey show that an average of 59% of the respondents agreed that the sessions met their expectations and reported the sessions as good, while 39 % of the respondents reported the outcome of the sessions as excellent. However, only 2 % reported the outcome of the sessions as fair. Therefore, this indicates that most respondents agreed that the training met their expectations. The relatively low number of participants who concluded that the sessions excellently met their expectations (39%) could be ascribed to the participants who joined online. The remote instructor-paced sessions may not have effectively engaged the learners, and the analysis of results and feedback suggests that remote self-paced learning could be improved in such training sessions in the future.

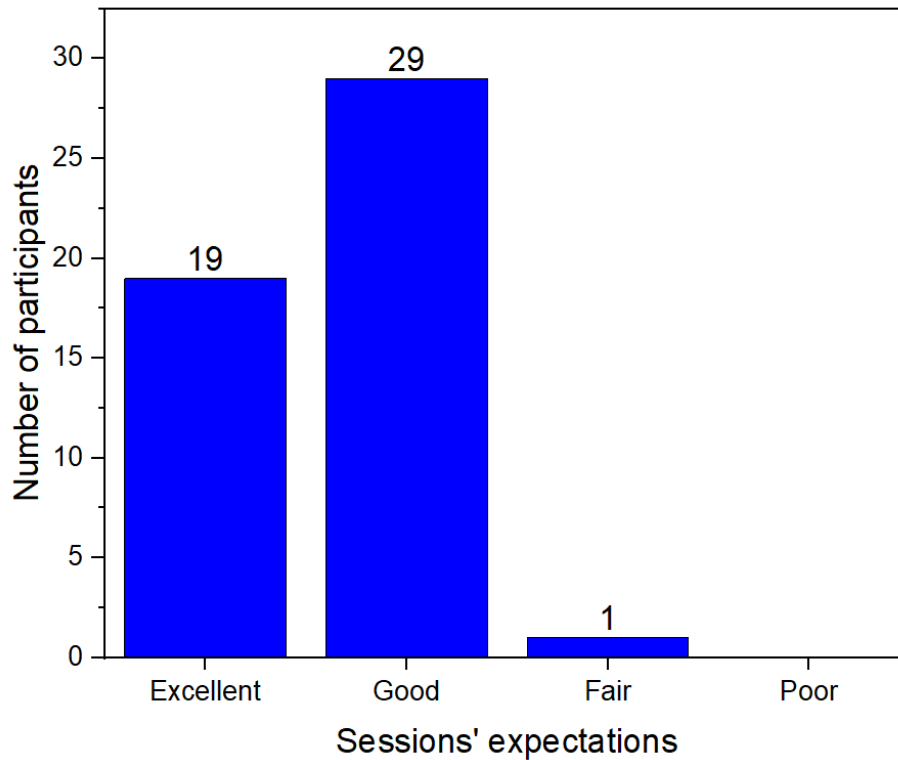


Figure 9: Respondents' feedback on the expectations from the sessions

3.1.8 Training evaluations

This training had a series of practical and open discussions that resulted in the overall training evaluation. The results from the training evaluations of the ML workshop are shown in Figure 10. From the results obtained after the survey, 45% and 55% of respondents reported the overall training evaluation as excellent and good, respectively. This indicates that the learners were satisfied with the approaches used for the workshop training.

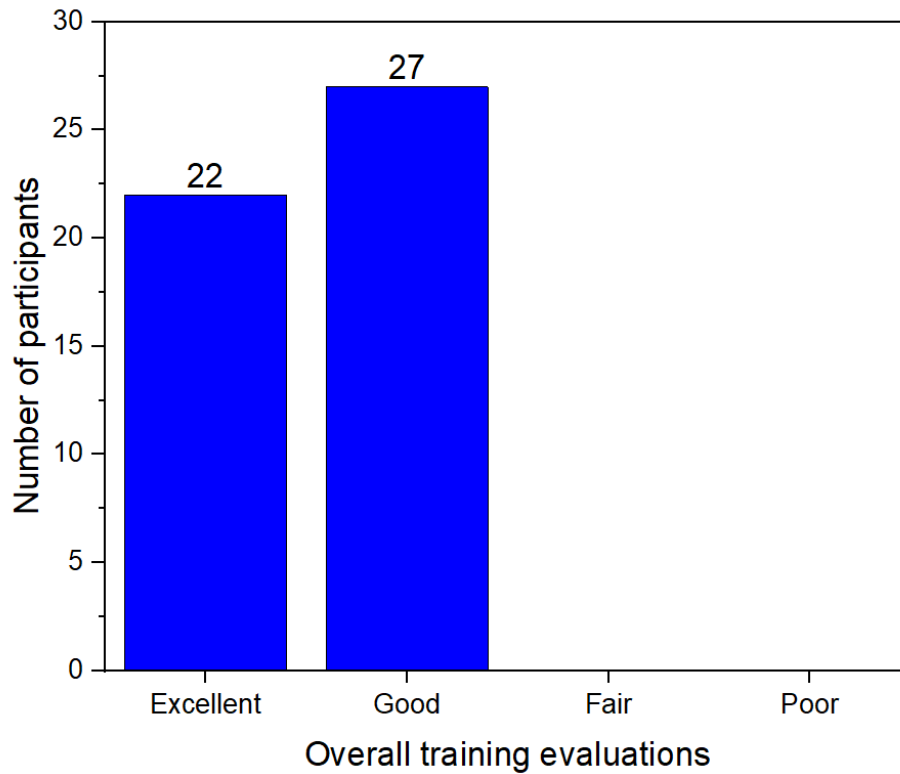


Figure 10: Respondents' feedback on the overall training evaluations

3.1.9 Recommendation of training sessions to their colleagues

The probability of the participants recommending the workshop to other colleagues was also evaluated based on the learner's satisfaction survey. From the results obtained as displayed in Figure 11, most of the respondents believe that they will recommend such training sessions to their colleagues. Their interest in recommending the training was rated from 0 to 10 with low to high levels of recommendation in ascending order. Out of 49 respondents, 45 % chose 10, which signifies the likelihood for them to recommend this training session to their colleagues, while others chose between 2 and 9. This indicates that the ML training was impactful and significant to the participants.

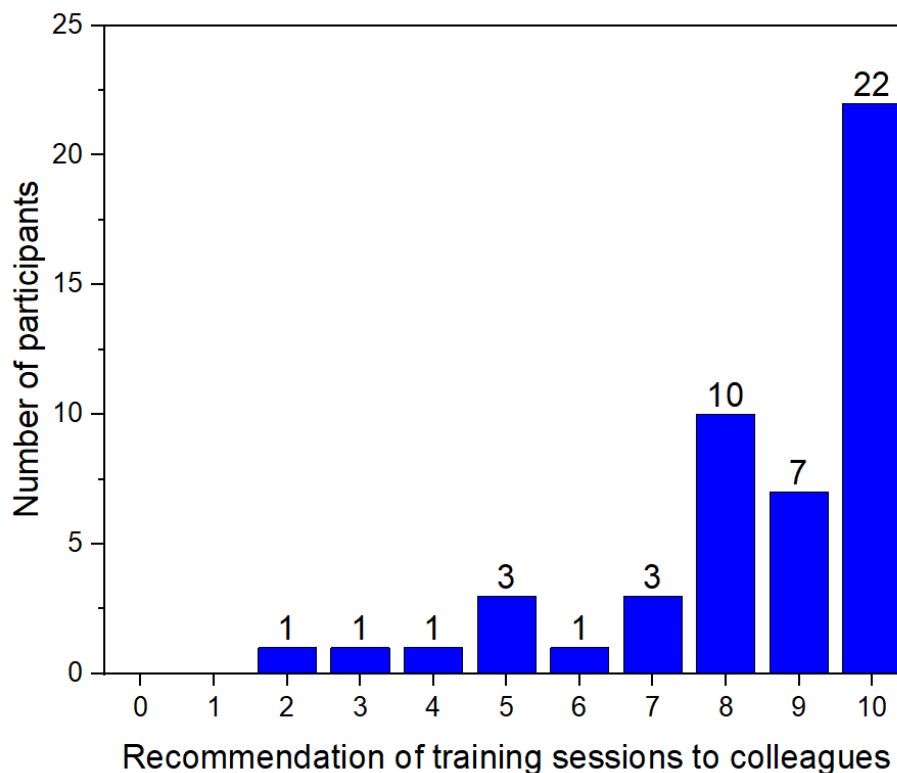


Figure 11: Respondents' feedback on recommending the training sessions to colleagues.

4. Conclusion

This paper outlines a workshop training session that focuses on teaching engineering students the fundamental principles of ML through a practical and interactive approach to address the disparity in the skill sets of engineering graduates in ML, a subsection of AI. The hands-on approach effectively enhanced the comprehension of the participants in understanding the ML concepts taught during the workshop. This was supported by a high level of expertise from the facilitators through the module contents, fascinating slides, and scripts uploaded on *Google Colab*. Also, the practical hands-on approach of the workshop bridged the gap between theory and practice. Additionally, a practical demonstration session exposed the students to the interpretation of results from ML tasks. Overall, the practical sessions improved the learner's understanding of the predictive capabilities of ML models in the realm of materials science. Results from the learner's satisfaction survey carried out after the training showed that a larger proportion of the participants were satisfied with the information obtained from the training, the facilitator's knowledge of the topics covered, the speaker's presentation skills, the content of slides and virtual aids, overall training evaluation, and recommendation of training sessions to colleagues to mention a few. It is recommended with the significant interest of the participants that ML can be integrated into curricula and research programs for institutions and organizations in the field of STEM and materials science. This will help to equip the next generation of students and researchers with the right knowledge and skills needed to harness the full potential of data-driven solutions in these disciplines.

Appendix A

A 1 Modules covered during the workshop.

This section consists of the modules and course contents developed by the facilitators to for the workshop. In what follows, the modules are described.

A 1.1 Introduction to ML Models

ML is a subset of Artificial Intelligence that employs computer algorithms to iteratively train and learn continuously on data without requiring explicit programming. There exist three primary categories of ML models: supervised, unsupervised, and reinforcement learning [29]. The participants were exposed to the supervised learning models. Datasets on the hydroxyapatite material as well as 199 cubic perovskite compounds were used to teach participants basic ML prediction tasks. An email containing the notebooks and materials was sent to participants, and they were also taught how to upload the documents on *Google Colab* for further analysis. The types of ML models were explained to the students on a fundamental level and a snapshot is highlighted.

A 1.1.1 Supervised Learning Models

This ML model operates by allowing the algorithms to learn from labeled data. It efficiently processes input data and accurately adjusts output labels by providing immediate feedback to predict the output. It is commonly employed to address regression or classification-based problems [33]. Supervised learning encompasses a variety of tasks, including regression (predicting numerical values), classification (predicting categorical values), and time series forecasting.

A 1.1.2 Unsupervised Learning Model

This ML model utilizes algorithms to analyze and construct a model from unlabelled data to uncover concealed structures and patterns within the data. Unsupervised learning models operate to reveal concealed insights and cluster similar data points together [33,34]. Unsupervised learning encompasses various techniques such as clustering, association, data reduction, and link prediction.

A 1.1.3 Reinforcement learning

This ML approach learns by trial and error through interaction with the environment to create optimal judgments that maximize cumulative rewards. This algorithm utilizes incentive systems to get feedback and acquire a sequence of actions [33,34]. Each of these learning models incorporates contemporary developmental methodologies that frequently combine one or more libraries and frameworks. Here are some covered in this workshop.

A 1.2 ML Frameworks

The participants were taught about ML frameworks. These are software tools or libraries that assist developers in constructing ML models or applications without requiring an extensive understanding of the underlying theories [35]. They offer a thorough process for the development of ML models. The selection of a framework relies on the nature of the application and the dataset being utilized. It is crucial to consider factors such as scalability, data processing, and deployment. Several popular ML frameworks include:

1. TensorFlow
2. Sci-Kit Learn
3. PyTorch

4. Keras.
5. Spark ML Lib etc.

The attendees were taught the Sci-Kit Learn ML framework.

A 1.2.1 Sci-Kit Learn

Sci-Kit Learn offers extensive assistance for ML experiments with its vast library specifically designed for the Python programming language. It is a renowned ML framework, widely recognized as an open-source Python tool for data mining and analysis. The software provides a diverse set of functionalities for developing algorithms and models in several fields, such as classification, clustering, pre-processing, regression, dimensional reduction, and model selection [20,35]. The main characteristics comprise:

- Effectively work with Python.
- It is regarded as the top framework for data mining and data analysis.
- It is free and open source.

To conduct ML experiments, it is necessary to have a specific set of features, which are also referred to as descriptors. Before making predictions, it is important to carefully identify the key features that are closely associated with the desired output. In this case, the approach employed for identifying features is feature engineering [36].

A 1.2.2 Feature Engineering

Feature engineering is a methodology employed in the field of ML, wherein data is utilized to generate new variables for the training set. This method can generate new features for both supervised and unsupervised learning tasks, aiming to streamline and expedite data manipulations while concurrently improving the precision of the models. It reduces the dimension of input space as much as possible without losing important information. Redundant and high self-correlation features are removed to guarantee the efficiency and accuracy of the models [37]. Processes involved in feature engineering include [32,37].

1. Correlation
2. Feature selection
3. Feature extraction

The correlation feature engineering process was utilized for this workshop. This method is frequently utilized at the initial stage of data preparation. In this approach, the selection of features is determined by the scores obtained from several statistical tests that measure their association with the result variable. For this workshop, the Pearson correlation plot was used to filter out the redundant features from the original datasets.

A 1.2.3 Pearson's Correlation

This assesses the degree of linear dependence between two variables, X and Y. The value of the variable ranges from a negative one (-1) to a positive one (+1) [31,38].

Pearson's correlation is given as:

$$r_{xy} = \frac{\sum(x_i - \bar{x}) \sum(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \sqrt{\sum(y_i - \bar{y})^2}}$$

Where $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ is the mean of y while $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ represent the mean of x . This shows how variables x and y are linearly highly associated with a positive correlation coefficient for directly related variables and negative for the ones that are inversely related.

Benefits of Pearson Correlation (r_{xy})

- Measures the similarity of two features ranging between -1 to 1.
- A value close to 1 indicates that two features have a high correlation and may be related.

Following an introductory session and practical experience on fundamental concepts such as different types of ML models, machine learning frameworks, and feature engineering, learners are now being introduced to various ensemble learning techniques.

A 1.3 Introduction to Ensemble Learning Techniques in ML.

Ensemble methods have been widely recognized in the field of data mining and ML in the past decade. They combine numerous models into a single entity, typically surpassing the accuracy of its components. The huge success of these ML ensemble learning techniques was documented in 1990 [39]. Various models may excel in different aspects of the data, even if they exhibit underfitting. Individual errors can be mitigated through averaging, particularly when models demonstrate overfitting. Bias-variance analysis informs us that we are faced with two choices: If the model exhibits underfitting (high bias, low variance), we can employ a technique known as Boosting, which involves a combination with other low-variance models. When a model exhibits overfitting, characterized by low bias and large variance, it can be mitigated by combining it with additional low-bias models using a technique known as Bagging. The models must exhibit no correlation, as any correlation will negatively impact the ensemble learning performance. Additionally, we can gain knowledge on the technique of merging the forecasts generated by various models through a process known as stacking [40].

Gradient boosting, Extreme Gradient (XGBoost), and Adaptive Boosting (AdaBoost) are instances of boosting algorithms, whereas Random Forest and extra trees classifiers are renowned bagging methods. Examples of the stacking framework include the utilization of super ensemble and blending techniques [40–42]. These ensemble methods have demonstrated exceptional performance in several ML applications. Their usage in numerous practical applications is likewise widely recognized [40]. Some of the ensemble learning techniques explained to the participants are:

A 1.3.1 Adaptive Boosting (AdaBoost)

This obtains different models by reweighting the training data at every iteration. It reduces underfitting by focusing on the 'hard' training examples. It increases the weights of instances misclassified by the ensemble, and vice versa. It is expected to be simple so that different instance weights lead to different models [40–43]. However, a limitation of AdaBoost is that it is sensitive to noisy data and outliers because of its iterative learning approach, causing overfitting.

A 1.3.2 Gradient Boosting

One of the key benefits of gradient boosting is its ability to learn intricate patterns from the input data by iteratively correcting the errors of the previous model, like other boosting algorithms. This model is highly regarded and extensively utilized. On the other hand, if the iterative task is not properly regularized, this algorithm may end up overfitting [40].

Furthermore, if the input data is noisy, there is a risk that a model built using this algorithm may overfit [44]. It excels in handling diverse features and varying scales. Generally, it outperforms the Random Forest ensemble learning technique, but it needs more fine-tuning and longer training. This algorithm is well-suited for applications with small datasets [45].

A 1.3.3 Extreme Gradient Boosting (XGBoost)

The XGBoost algorithm utilizes the gradient boosting framework and is based on decision trees, making it a powerful ensemble method. This algorithm is known for its scalability and high accuracy in classification and regression applications [40]. Like gradient boosting, it utilizes various techniques such as controlling tree depth, adjusting the learning rate, and subsampling to effectively address the issue of overfitting. This algorithm has the advantage of requiring minimal feature engineering, such as data normalization and feature scaling, as it can handle these situations effectively. In addition, it can handle missing values. Also, it can provide feature importance, allowing for a deeper understanding of the input features and enabling feature selection. It is known for its speed, ability to handle large datasets, and resistance to overfitting [46]. Although the model is well-crafted, it does have a few limitations, such as a high number of hyperparameters that can be challenging to fine-tune [47].

A 1.3.4 LightGBM

This technique utilizes gradient-based sampling to rapidly enhance model performance. It has applications in classification, ranking, and other ML tasks. The LightGBM algorithm employs two innovative techniques, Gradient-based One-Sided Sampling (GOSS) and Exclusive Feature Bundling (EFB), to enhance training speed and produce superior accuracy. The GOSS technique is a variant of the gradient boosting technique that prioritizes training instances with greater gradients. This approach accelerates the learning process and decreases the computational complexity of the model. The rationale for removing samples with modest gradients is that occurrences with significant gradients are more valuable in determining the information gain [40]. The EFB technique reduces the number of characteristics by bundling sparse mutually exclusive attributes, thus performing a feature selection task [48]. LightGBM offers a notable advantage in terms of speed and consistently produces highly efficient models. Additionally, it exhibits minimal memory usage as it transforms continuous values into discrete bins. Furthermore, it attains significantly superior precision compared to other boosting strategies because of incorporating the GOSS and EFB procedures. Furthermore, the LightGBM method exhibits superior performance when trained on extensive datasets, boasting a quicker training time compared to the XGBoost algorithm [49]. One drawback of LightGBM is its tendency to overfit short training datasets, while it is optimized for larger datasets. Furthermore, dividing the tree based on individual leaves may lead to overfitting as it generates more intricate trees.

A 1.3.5 CatBoost

This technique efficiently manages categorical features throughout the training phase. CatBoost has made a significant advancement by incorporating unbiased gradient estimates, which effectively mitigates the problem of overfitting. Thus, to calculate the slope of each instance at each boosting iteration, the CatBoost method excludes that instance from being utilized to train the current model [50]. Furthermore, a noteworthy enhancement in the CatBoost algorithm is its automatic conversion of category information into numerical representations. Categorical characteristics consist of a finite collection of values referred to as categories, which generally cannot be compared. Therefore, these characteristics are currently unsuitable for constructing decision trees. During the preprocessing stage, categorical features are frequently transformed into numerical features by substituting them with numerical values.

It has exceptional performance and surpasses most ML algorithms in scenarios where the input consists of categorical data. Additionally, it possesses the intrinsic capability to effectively manage missing data. Nevertheless, the performance of the system may be subpar if the parameters are not adequately fine-tuned [40].

During ML experiments, the performance of all the algorithms mentioned above can be accessed through performance metrics. This is utilized to compare the predictions made by the trained model with the actual (observed) data obtained from the testing data set [51]. The outcomes of these comparisons can directly impact the decision-making process of choosing the specific ML algorithms for deployment. Furthermore, these modules were instructed during the workshop session.

A 1.4 Performance Metrics in ML

Performance evaluations are indispensable in ML. The system assesses ensemble learning techniques and provides feedback on their efficacy. ML activities can be categorized as either regression or classification tasks. These tasks require performance metrics to assess their efficiency and comprehend how different models interpret data. Typically, the performance of a model is assessed by calculating the discrepancy between predicted values and actual observations using various statistical techniques [52]. Various metrics are employed to assess the performance of ML models in regression tasks, such as the Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), and R-squared (R^2). Their formulas are shown in equations (1) - (4) below:

$$MAE = \frac{\sum_{i=1}^N |y_i - \hat{y}_i|}{N} \text{----- (1)}$$

$$MSE = \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N} \text{----- (2)}$$

$$R^2 = 1 - \frac{\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N}}{\frac{\sum_{i=1}^N (y_i - \bar{y})^2}{N}} \text{----- (3)}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N}} \text{----- (4)}$$

It is assumed that the dataset contains N samples, \hat{y}_i is the predicted value for the i-th data point, y_i is the known value for the i-th data point [28].

A 1.4.1 Mean Absolute Error (MAE)

MAE is calculated as the average of the absolute differences between the actual values and the anticipated values. Mathematically, it is denoted by equation (1). It provides quantification of the discrepancy between the forecasts and the actual result. Nevertheless, MAE fails to provide insight into the direction of the error, specifically when the models are under-predicting or over-predicting the data.

A 1.4.2 Mean Squared Error (MSE)

The utilization of MSE depicted in equation (2) for statistical regression is extensively widespread. The utilization of it in the model leads to an overestimation of its performance by amplifying even slight errors. The inclusion of the squaring component makes it intrinsically more vulnerable to the impact of outliers in comparison to other metrics. This can be easily accomplished by utilizing NumPy arrays in the Python programming language.

A 1.4.3 R² Coefficient of determination

R² is a post-metric measure, which implies that it is derived from other metrics. The coefficient in question quantifies the extent to which the variability in the dependent variable Y (goal) may be attributed to the variability in the independent variable X (regression line), expressed as a percentage. This coefficient is represented by equation 3.

A few intuitions related to R² results are:

- If the sum of squared errors of the regression line is modest, it indicates that the coefficient of determination (R²) will be near 1, which is considered ideal. This implies that the regression model has successfully captured 100% of the variance in the target variable.
- On the other hand, in cases where the sum of the squared error of the regression line is considerable, the coefficient of determination (R²) will approach 0, indicating that the regression model failed to explain the variability observed in the dependent variable.

A 1.4.4 Root Mean Squared Error (RMSE)

RMSE is a statistical measure used to assess the accuracy of a prediction model. It is calculated by taking the square root of the average of the squared differences between the predicted values and the actual values. It maintains the differentiable characteristic MSE and mitigates the errors of MSE by applying the square root function to it (equation (4)). The interpretation of errors can be easily accomplished as the scale now matches that of the random variable. Due to the normalization of the scale components, data is less susceptible to outliers.

Establishing the importance of the performance metric of each ensemble learning technique is essential and highly significant as it unveils their efficiencies. On the other hand, the contribution of each of the features used in experimenting with ensemble learning techniques cannot be over-emphasized. This can be made visible through SHapley Additive exPlanations (SHAP) which allows the visualization of each input feature according to their rank and contribution during the experiment. Attendees were exposed to this tool during the workshop session.

A 1.5 SHapley Additive exPlanations (SHAP) in ML

To provide a theoretical explanation of the model rationale behind the superiority of one feature over another. SHAP algorithm assesses the impact of the feature importance when the supervised learning algorithms are used in experimenting [28].

This is a primer on explaining ML models using Shapley values. It is a popular cooperative game theory strategy with attractive qualities. Within the field of ML, it is typical to assign a significance measure to each feature to quantify its respective contribution towards the overall output of the model. The SHAP values provide insights into the impact of individual features on the final prediction, the relative importance of each feature compared to others, and the extent to which the model depends on the interaction among features. This method provides a

good understanding of how complex models make decisions and the importance of model interpretability.

SHAP analyses are model agnostic, which means they can be utilized to interpret any ML, such as Decision trees, Random forests, Gradient boosting models, Neural networks etc.

This tool provides various beneficial characteristics that render them effective for model interpretation. These are:

- **Additivity:** - SHAP analysis are additive, therefore each feature prediction contribution can be calculated separately and added. This makes SHAP calculation efficient even for high-dimensional datasets.
- **Local accuracy:** - It finds the difference between the expected model output and the actual output for a given input. Therefore, SHAP accurately and locally interprets the model's prediction for a particular input.
- **Consistency:** - Model modifications do not affect SHAP unless a feature contributes. SHAP interprets model behaviour consistently regardless of model design or parameters.

In general, SHAP offers a reliable and unbiased approach to acquiring an understanding of the prediction process of an ML model and identifying the features that exert the most significant impact.

Based on the variation observed in the feature ranking of various ensemble learning models using SHAP. A novel holistic feature ranking method was initiated by the facilitators to determine the global feature ranking across all the explained models. This novel insight was also imparted to the participants.

A 1.6 Novel Holistic Feature Ranking Method


This method has been developed to identify input features that contribute the most to the predictions of the trained supervised learning model and is represented by equation (5). Details of the method have been reported in the work of Obada et al., [28].

$$R_s(F_i^p) = \sum_i count(F_i^p) \times p \text{ --- (5)}$$

Appendix B

B 1 A Snippet of the slides used during the Workshop.

B 1. 1 Introduction to Machine Learning Frameworks.

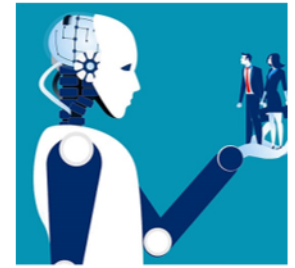


Introduction to Machine Learning Frameworks


Basic Concepts in Machine Learning for Materials Discovery, October 6th to 11th, 2023.

Machine learning tasks

- **Supervised learning**
 - regression: predict numerical values
 - classification: predict categorical values, i.e., labels
- **Unsupervised learning**
 - clustering: group data according to "distance"
 - association: find frequent co-occurrences
 - link prediction: discover relationships in data
 - data reduction: project features to fewer features
- **Reinforcement learning**



Applications



Materials Science in Semiconductor Processing

Volume 101, July 2023, 107507

Explainable machine learning for predicting the band gaps of ABX_3 perovskites

Barak D. Ghosh,^{1,2,3} A. W. Barnard,^{1,2,3} J. W. Johnson,⁴ M. J. Heule,^{1,2,3} M. J. Heule,^{1,2,3} D. J. W. Simons,^{1,2,3} M. J. Heule,^{1,2,3} A. W. Barnard,^{1,2,3}



Machine learning for molecular and materials science

Biorganic & Medicinal Chemistry Letters

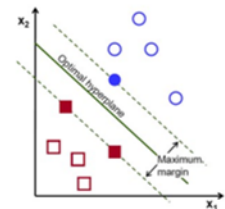
Recent applications of machine learning in medicinal chemistry

Machine Learning Applications for Earth Observation

The Application of Machine Learning in Biology

Machine learning algorithms

- **Regression:** Ridge regression, Support Vector Machines, Random Forest, Multilayer Neural Networks, Deep Neural Networks, ...
- **Classification:** Naive Base, Support Vector Machines, Random Forest, Multilayer Neural Networks, Deep Neural Networks, ...



Frameworks

• Programming languages

- Python
- R
- C++
- ...

• Many libraries

- scikit-learn
- PyTorch
- TensorFlow
- Keras
- ...



classic machine learning

deep learning frameworks

scikit-learn

- **Nice end-to-end framework**
 - data exploration (+ pandas + holoviews)
 - data preprocessing (+ pandas)
 - cleaning/missing values
 - normalization
 - training
 - testing
 - application
- "Classic" machine learning only
- <https://scikit-learn.org/stable/>



B 1. 2 Performance Metrics for ML Techniques.

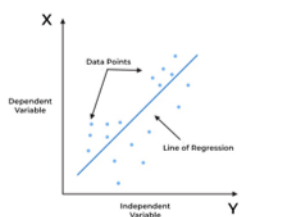


Performance Metrics Used for Machine Learning Techniques

Basic Concepts in Machine Learning for Materials Discovery, October 6th to 11th, 2023

Machine Learning Performance Metrics

- **Classification**
 - Confusion Matrix
 - Accuracy
 - Recall or Sensitivity
 - Precision
 - F1 Score
- **Regression**
 - Mean Absolute Error
 - Mean Square Error

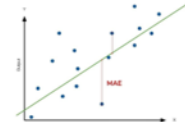


Regression

Mean Absolute Error

Mean Absolute Error is the average of the difference between the original values and the predicted values

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$



Mean Squared Error

Mean Squared Error is the average of the square difference between the original values and the predicted values

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$



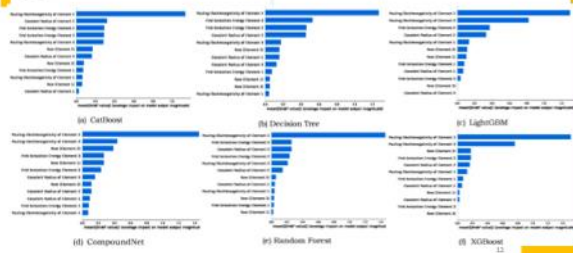
B 1. 3 The Novel Holistic Ranking Method.



The Novel Holistic Ranking Method

Basic Concepts in Machine Learning for Materials Discovery, October 6th to 11th, 2023

Results SHAP Explainability Assessment



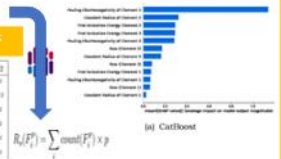
Holistic Ranking

Table 4: Explainability of the model feature importance ranking in the testing phase; the power index represents the feature positional ranking

Methods	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
CATBOOST	1 st	2 nd	7 th	3 rd	2 nd	10 th	11 th	6 th	9 th	8 th	5 th	12 th
XGBOOST	1 st	2 nd	10 th	7 th	3 rd	2 nd	11 th	6 th	9 th	8 th	5 th	12 th
RANDOM FOREST	1 st	2 nd	3 rd	7 th	4 th	5 th	10 th	6 th	9 th	8 th	11 th	12 th
COMPOUNDNET	1 st	2 nd	10 th	7 th	3 rd	2 nd	11 th	6 th	9 th	8 th	5 th	12 th
LIGHTGBM	1 st	2 nd	7 th	3 rd	2 nd	10 th	11 th	6 th	9 th	8 th	5 th	12 th
DECISION TREE	1 st	2 nd	7 th	3 rd	2 nd	10 th	11 th	6 th	9 th	8 th	5 th	12 th

- R_e - sum of effective ranking performance of the learning models
- F_i^p - the data input features
- p - positional ranking of features in each of the explained machine learning models

Components	PRE.1	PRE.2	PRE.3	ONE.1	ONE.2	PRE.1	PRE.2	PRE.3	Rand.1	Rand.2	Rand.3	
PATCH	1.07	1.34	3.44	325	128	75	102.8	68.8	333.9	6	4	2
RAMOD	0.88	1.40	3.44	308	127	75	102.9	69.1	333.9	6	5	2
TZDF	1.02	1.05	3.08	348	121	75	108.4	80.4	303.9	6	4	2



Holistic Ranking Cont'd

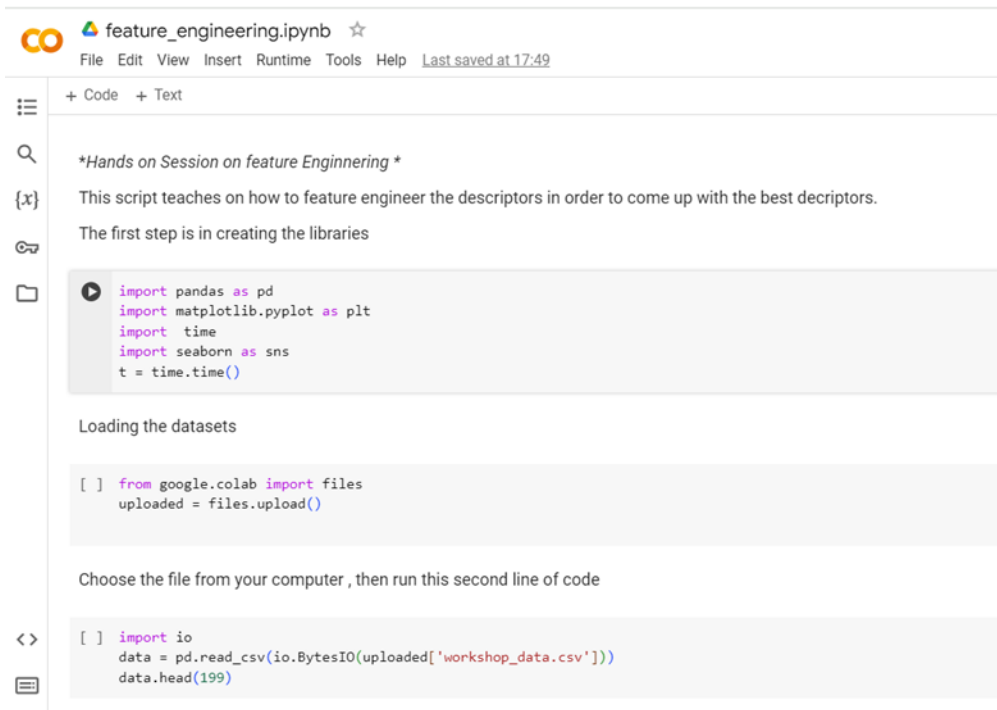
Table 5: Holistic feature ranking using all the outcomes from the explained machine learning models

Features	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
Sum of ranking R_e	6	21	23	30	31	33	47	54	54	54	55	60
Best feature ranking order	1	2	7	4	10	8	5	9	3	9	6	11
Feature importance	Pulling Electro-negativity of Element 1	Pulling Electro-negativity of Element 2	First Ionization Energy of Element 2	Constant Radius of Element 2	Row (Element 2)	First Ionization Energy Element 3	Constant Radius of Element 3	Pulling Electro-negativity of Element 1	Constant Radius of Element 1	Row (Element 1)	First Ionization Energy Element 1	Row (Element 3)

Appendix C

C 1 A snippet of ipynb files used during the Workshop.

C 1.1 A Snippet of ipynb files for feature engineering.



The screenshot shows a Jupyter Notebook interface for a file named 'feature_engineering.ipynb'. The notebook contains the following content:

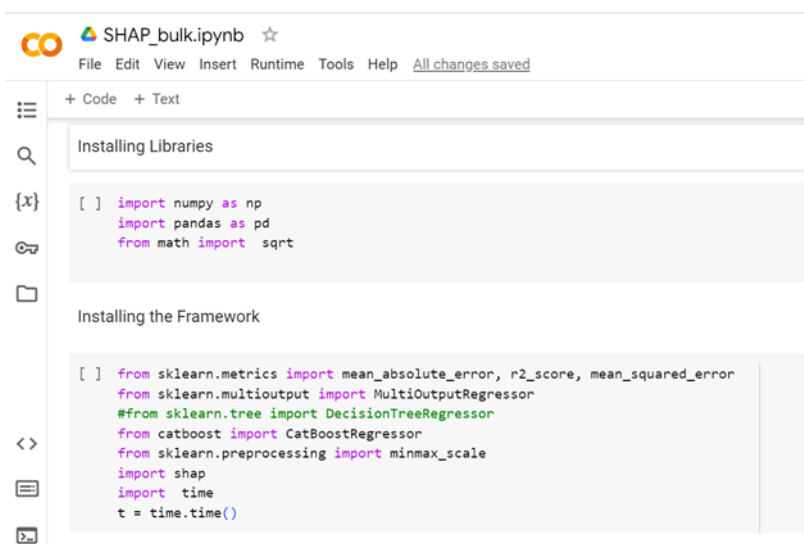
- A search bar with the text: **Hands on Session on feature Engineering**
- A text cell: This script teaches on how to feature engineer the descriptors in order to come up with the best descriptors. The first step is in creating the libraries
- A code cell with the following Python code:

```
import pandas as pd
import matplotlib.pyplot as plt
import time
import seaborn as sns
t = time.time()
```
- A text cell: Loading the datasets
- A code cell with the following Python code:

```
from google.colab import files
uploaded = files.upload()
```
- A text cell: Choose the file from your computer , then run this second line of code
- A code cell with the following Python code:

```
import io
data = pd.read_csv(io.BytesIO(uploaded['workshop_data.csv']))
data.head(199)
```

C 1. 2 A Snippet of ipynb files for SHAP Analysis.



The screenshot shows a Jupyter Notebook interface for a file named 'SHAP_bulk.ipynb'. The notebook contains the following content:

- A search bar with the text: Installing Libraries
- A code cell with the following Python code:

```
import numpy as np
import pandas as pd
from math import sqrt
```
- A text cell: Installing the Framework
- A code cell with the following Python code:

```
from sklearn.metrics import mean_absolute_error, r2_score, mean_squared_error
from sklearn.multioutput import MultiOutputRegressor
#from sklearn.tree import DecisionTreeRegressor
from catboost import CatBoostRegressor
from sklearn.preprocessing import minmax_scale
import shap
import time
t = time.time()
```


References

- [1] W. Rawat, Z. Wang, Deep convolutional neural networks for image classification: A comprehensive review, *Neural Computation* 29 (2017) 2352–2449.
- [2] H. Sak, A.W. Senior, F. Beaufays, Long short-term memory recurrent neural network architectures for large scale acoustic modeling, (2014). https://www.isca-archive.org/interspeech_2014/sak14_interspeech.pdf (accessed April 1, 2024).
- [3] Y. Liu, J. Zhang, Deep Learning in Machine Translation, in: L. Deng, Y. Liu (Eds.), *Deep Learning in Natural Language Processing*, Springer Singapore, Singapore, 2018: pp. 147–183. https://doi.org/10.1007/978-981-10-5209-5_6.
- [4] A. Theissler, Detecting known and unknown faults in automotive systems using ensemble-based anomaly detection, *Knowledge-Based Systems* 123 (2017) 163–173.
- [5] A. Theissler, J. Pérez-Velázquez, M. Kettelgerdes, G. Elger, Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry, *Reliability Engineering & System Safety* 215 (2021) 107864.
- [6] T. Markert, S. Matich, E. Hoerner, A. Theissler, M. Atzmueller, Fingertip 6-axis force/torque sensing for texture recognition in robotic manipulation, in: *2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, IEEE, 2021: pp. 1–8. <https://ieeexplore.ieee.org/abstract/document/9613688/> (accessed April 1, 2024).
- [7] M. Gaertner, A. Theissler, M. Fernandes, Detecting Potential Subscribers on Twitch: A Text Mining Approach with XGBoost - Discovery Challenge ChAT: CoolStoryBob, 2020.
- [8] A. Theissler, P. Ritzer, EduML: An explorative approach for students and lecturers in machine learning courses, in: *2022 IEEE Global Engineering Education Conference (EDUCON)*, IEEE, 2022: pp. 921–928. .
- [9] D. Touretzky, C. Gardner-McCune, F. Martin, D. Seehorn, Envisioning AI for K-12: What Should Every Child Know about AI?, *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (2019) 9795–9799.
- [10] L. Wunderlich, A. Higgins, Y. Lichtenstein, Machine Learning for Business Students: An Experiential Learning Approach, in: *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V. 1*, ACM, Virtual Event Germany, 2021: pp. 512–518. <https://doi.org/10.1145/3430665.3456326>.
- [11] D. Huppenkothen, G. Eadie, Teaching the foundations of machine learning with candy, in: *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, PMLR, 2021: pp. 29–35.
- [12] Cc2020 Task Force, *Computing Curricula 2020: Paradigms for Global Computing Education*, ACM, New York, NY, USA, 2020. <https://doi.org/10.1145/3467967>.
- [13] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi, T. Zimmermann, Software Engineering for Machine Learning: A Case Study, in: *2019 IEEE/ACM 41st International Conference on Software Engineering: Software*

- Engineering in Practice (ICSE-SEIP), IEEE, Montreal, QC, Canada, 2019: pp. 291–300. <https://doi.org/10.1109/ICSE-SEIP.2019.00042>.
- [14] L.S. Marques, C. Gresse von Wangenheim, J.C. Hauck, Teaching machine learning in school: A systematic mapping of the state of the art, *Informatics in Education* 19 (2020) 283–321.
- [15] R.M. Martins, C. Gresse Von Wangenheim, Findings on teaching machine learning in high school: A ten-year systematic literature review, *Informatics in Education* 22 (2023) 421–440.
- [16] E. Sulmont, E. Patitsas, J.R. Cooperstock, What Is Hard about Teaching Machine Learning to Non-Majors? Insights from Classifying Instructors' Learning Goals, *ACM Trans. Comput. Educ.* 19 (2019) 1–16. <https://doi.org/10.1145/3336124>.
- [17] K. Mike, T. Hazan, O. Hazzan, Equalizing Data Science Curriculum for Computer Science Pupils, in: *Koli Calling '20: Proceedings of the 20th Koli Calling International Conference on Computing Education Research*, ACM, Koli Finland, 2020: pp. 1–5. <https://doi.org/10.1145/3428029.3428045>.
- [18] A.A. Reyes, C. Elkin, Q. Niyaz, X. Yang, S. Paheding, V.K. Devabhaktuni, A Preliminary Work on Visualization-based Education Tool for High School Machine Learning Education, in: *2020 IEEE Integrated STEM Education Conference (ISEC)*, IEEE, Princeton, NJ, USA, 2020: pp. 1–5.
- [19] L. Huang, K.-S. Ma, Introducing Machine Learning to First-year Undergraduate Engineering Students Through an Authentic and Active Learning Labware, in: *2018 IEEE Frontiers in Education Conference (FIE)*, IEEE, San Jose, CA, USA, 2018: pp. 1–4. <https://doi.org/10.1109/FIE.2018.8659308>.
- [20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: Machine learning in Python, *The Journal of Machine Learning Research* 12 (2011) 2825–2830.
- [21] S. Kim, E.C. Bucholtz, K. Briney, A.P. Cornell, J. Cuadros, K.D. Fulfer, T. Gupta, E. Hepler-Smith, D.H. Johnston, A.S.I.D. Lang, D. Larsen, Y. Li, L.R. McEwen, L.A. Morsch, J.L. Muzyka, R.E. Belford, Teaching Cheminformatics through a Collaborative Intercollegiate Online Chemistry Course (OLCC), *J. Chem. Educ.* 98 (2021) 416–425. <https://doi.org/10.1021/acs.jchemed.0c01035>.
- [22] J. Perna, Possibilities and Challenges of Using Educational Cheminformatics for STEM Education: A SWOT Analysis of a Molecular Visualization Engineering Project, *J. Chem. Educ.* 99 (2022) 1190–1200. <https://doi.org/10.1021/acs.jchemed.1c00683>.
- [23] E. Chen, M. Asta, Using Jupyter Tools to Design an Interactive Textbook to Guide Undergraduate Research in Materials Informatics, *J. Chem. Educ.* 99 (2022) 3601–3606. <https://doi.org/10.1021/acs.jchemed.2c00640>.
- [24] J.M. Perkel, Programming: Pick up python, *Nature* 518 (2015) 125–126.
- [25] M. Van Staveren, Integrating Python into a Physical Chemistry Lab, *J. Chem. Educ.* 99 (2022) 2604–2609. <https://doi.org/10.1021/acs.jchemed.2c00193>.
- [26] C.J. Weiss, A Creative Commons Textbook for Teaching Scientific Computing to Chemistry Students with Python and Jupyter Notebooks, *J. Chem. Educ.* 98 (2021) 489–494. <https://doi.org/10.1021/acs.jchemed.0c01071>.

- [27] W. Vallejo, C. Díaz-Urbe, C. Fajardo, Google Colab and Virtual Simulations: Practical e-Learning Tools to Support the Teaching of Thermodynamics and to Introduce Coding to Students, *ACS Omega* 7 (2022) 7421–7429.
- [28] D.O. Obada, E. Okafor, S.A. Abolade, A.M. Ukpong, D. Dodoo-Arhin, A. Akande, Explainable machine learning for predicting the band gaps of ABX_3 perovskites, *Materials Science in Semiconductor Processing* 161 (2023) 107427. <https://doi.org/10.1016/j.mssp.2023.107427>.
- [29] L. Fiedler, K. Shah, M. Bussmann, A. Cangi, Deep dive into machine learning density functional theory for materials science and chemistry, *Phys. Rev. Materials* 6 (2022) 040301. <https://doi.org/10.1103/PhysRevMaterials.6.040301>.
- [30] S. Körbel, M.A.L. Marques, S. Botti, Stability and electronic properties of new inorganic perovskites from high-throughput ab initio calculations, *J. Mater. Chem. C* 4 (2016) 3157–3167. <https://doi.org/10.1039/C5TC04172D>.
- [31] J. Benesty, J. Chen, Y. Huang, I. Cohen, Pearson Correlation Coefficient, in: *Noise Reduction in Speech Processing*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009: pp. 1–4. https://doi.org/10.1007/978-3-642-00296-0_5.
- [32] S. Jain, A. Saha, Rank-based univariate feature selection methods on machine learning classifiers for code smell detection, *Evol. Intel.* 15 (2022) 609–638. <https://doi.org/10.1007/s12065-020-00536-z>.
- [33] T. Talaei Khoei, N. Kaabouch, Machine Learning: Models, Challenges, and Research Directions, *Future Internet* 15 (2023) 332.
- [34] T. Hastie, J. Friedman, R. Tibshirani, *The Elements of Statistical Learning*, Springer New York, New York, NY, 2001. <https://doi.org/10.1007/978-0-387-21606-5>.
- [35] G. Nguyen, S. Dlugolinsky, M. Bobák, V. Tran, Á. López García, I. Heredia, P. Malík, L. Hluchý, Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey, *Artif Intell Rev* 52 (2019) 77–124. <https://doi.org/10.1007/s10462-018-09679-z>.
- [36] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, Machine learning in materials informatics: recent applications and prospects, *Npj Computational Materials* 3 (2017) 54.
- [37] Q. Tao, P. Xu, M. Li, W. Lu, Machine learning for perovskite materials design and discovery, *Npj Computational Materials* 7 (2021) 23.
- [38] F. Sustainability, Analyzing meteorological parameters using Pearson correlation coefficient and implementing machine learning models for solar energy prediction in Kuching, Sarawak | *Future Sustainability*, (2024).
- [39] T.N. Rincy, R. Gupta, Ensemble Learning Techniques and its Efficiency in Machine Learning: A Survey, in: *2nd International Conference on Data, Engineering and Applications (IDEA)*, IEEE, Bhopal, India, 2020: pp. 1–6.
- [40] I.D. Mienye, Y. Sun, A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects, *IEEE Access* 10 (2022) 99129–99149.

- [41] I.K. Nti, A.F. Adekoya, B.A. Weyori, A comprehensive evaluation of ensemble learning for stock-market prediction, *J Big Data* 7 (2020) 20. <https://doi.org/10.1186/s40537-020-00299-5>.
- [42] M.H.D.M. Ribeiro, L. dos Santos Coelho, Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series, *Applied Soft Computing* 86 (2020) 105837.
- [43] F. Wang, Z. Li, F. He, R. Wang, W. Yu, F. Nie, Feature Learning Viewpoint of Adaboost and a New Algorithm, *IEEE Access* 7 (2019) 149890–149899.
- [44] B. Zhang, J. Ren, Y. Cheng, B. Wang, Z. Wei, Health Data Driven on Continuous Blood Pressure Prediction Based on Gradient Boosting Decision Tree Algorithm, *IEEE Access* 7 (2019) 32423–32433.
- [45] J. Jiang, R. Wang, M. Wang, K. Gao, D.D. Nguyen, G.-W. Wei, Boosting Tree-Assisted Multitask Deep Learning for Small Scientific Datasets, *J Chem Inf Model* 60 (2020) 1235–1244. <https://doi.org/10.1021/acs.jcim.9b01184>.
- [46] W. Liang, S. Luo, G. Zhao, H. Wu, Predicting Hard Rock Pillar Stability Using GBDT, XGBoost, and LightGBM Algorithms, *Mathematics* 8 (2020) 765.
- [47] B. Zhang, Y. Zhang, X. Jiang, Feature selection for global tropospheric ozone prediction based on the BO-XGBoost-RFE algorithm, *Sci Rep* 12 (2022) 9244.
- [48] K. Neshatian, L. Varn, Feature Bundles and their Effect on the Performance of Tree-based Evolutionary Classification and Feature Selection Algorithms, in: *2019 IEEE Congress on Evolutionary Computation (CEC)*, IEEE, Wellington, New Zealand, 2019: pp. 1612–1619.
- [49] D.A. McCarty, H.W. Kim, H.K. Lee, Evaluation of Light Gradient Boosted Machine Learning Technique in Large Scale Land Use and Land Cover Classification, *Environments* 7 (2020) 84.
- [50] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, M. Asadpour, Boosting methods for multi-class imbalanced data classification: an experimental review, *J Big Data* 7 (2020) 70. <https://doi.org/10.1186/s40537-020-00349-y>.
- [51] A. Botchkarev, A New Typology Design of Performance Metrics to Measure Errors in Machine Learning Regression Algorithms, *IJKM* 14 (2019) 045–076. <https://doi.org/10.28945/4184>.
- [52] C.A. Gueymard, A review of validation methodologies and statistical performance indicators for modeled solar radiation data: Towards a better bankability of solar projects, *Renewable and Sustainable Energy Reviews* 39 (2014) 1024–1034. <https://doi.org/10.1016/j.rser.2014.07.117>.