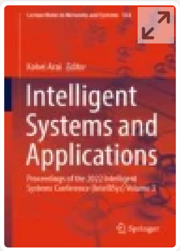


Proceedings of SAI Intelligent Systems Conference

IntelliSys 2022: **Intelligent Systems and Applications** pp 1–15

How to Improve the Teaching of Computational Machine Learning Applied to Large-Scale Data Science: The Case of Public Universities in Mexico

Sergio Rogelio Tinoco-Martínez, Heberto Ferreira-Medina, José Luis Cendejas-Valdez, Froylan Hernández-Rendón, Mariana Michell Flores-Monroy & Bruce Hiram Ginori-Rodríguez

Conference paper | First Online: 01 September 2022

Part of the Lecture Notes in Networks and Systems book series (LNNS,volume 544)

Abstract

Teaching along with training on Machine Learning (ML) and Big Data in Mexican universities has become a necessity that requires the application of courses, handbooks, and practices that allow improvement in the learning of Data Science (DS) and Artificial Intelligence (AI) subjects. This work shows how the academy and the Information Technology industry use tools to analyze large volumes of data to support decision-making, which is hard to treat and interpret directly. A solution to some large-scale national problems is the inclusion of these subjects in related courses within specialization areas that universities offer. The methodology in this work is as follows: 1) Selection of topics and tools for ML and Big Data teaching, 2) Design of practices with application to real data problems, and 3) Implementation and/or application of these practices in a specialization diploma. Results of a survey applied to academic staff and students are shown. The survey respondents have already taken related courses along with those specific topics that the proposed courses and practices will seek to strengthen, developing needed skills for solving problems where ML/DL and Big Data are an outstanding alternative of solution.

Keywords

Machine learning **Deep learning** **Big data** **Data science**

Teaching skills

References

1. Haenlein, M., Kaplan, A.: A brief history of artificial intelligence: on the past, present, and future of artificial intelligence. Calif. Manage. Rev. **61**(4), 5–14 (2019)

2. ENES-UNAM Homepage: http://www.enesmorelia.unam.mx/. Last accessed 18 Jan 2021

3. Nilsson, N.J.: Introduction to Machine Learning. Not published, Stanford, CA (1996)

4. Rumelhart, D., Hinton, G., Williams, R.: Learning representations by back-propagating errors. Nature **323**, 533–536 (1986)

5. Deng, L., Yu, D.: Deep learning: methods and applications. FNT Sign. Process. **7**, 197–387 (2014)

6. PyTorch Homepage: https://pytorch.org/. Last accessed 28 Jan 2021

7. Fast.ai Homepage: https://www.fast.ai/. Last accessed 28 Jan 2021

8. TensorFlow Homepage: https://www.tensorflow.org/. Last accessed 28 Jan 2021

9. Keras Homepage: https://keras.io/. Last accessed 28 Jan 2021

10. Deeplearning4j Homepage: https://deeplearning4j.konduit.ai. Last accessed 28 Jan 2021

11. El Naqa, I., Murphy, M.J.: What is machine learning? In: El Naqa, I., Li, R., Murphy, M.J. (eds.) Machine Learning in Radiation Oncology, pp. 3–11. Springer International Publishing, Cham (2015). https://doi.org/10.1007/978-3-319-18305-3 1

12. Géron, A.: Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media (2019)

13. Hagerty, J., Stanley R., Stoecker W.: Medical image processing in the age of deep learning. In: Proceedings of the 12th international joint conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP), pp. 306–311 (2017)

14. Basheer, I., Hajmeer, M.: Artificial neural networks: fundamentals, computing, design, and application. J. Microbiol. Methods **43**(1), 3–31 (2000)

15. Lévy J., Flórez R., Rodríguez J.: Las redes neuronales artificiales, 1a. Edición. Tirant lo Blanch., Netbiblo (2008)

16. Lasi, H., Fettke, P., Feld, T., Hoffmann, M.: Industry 4.0. Bus. Inform. Syst. Eng. **6**(4), 239–242 (2014)

17. Jordan, M., Mitchell, T.: Machine Learning: Trends, Perspectives, and Prospects. Science **349**(6245), 255–260 (2015)

18. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of Machine Learning. MIT Press (2018)

19. Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., Li, B.: Manipulating machine learning: poisoning attacks and countermeasures for regression learning. In: 2018 IEEE Symposium on Security and Privacy (SP), pp. 19–35 (2018)

20. Li, W., Han, J., Pei, J.: CMAR: Accurate and efficient classification based on multiple class-association rules. In: Proceedings 2001 IEEE. International Conference on Data Mining, pp. 369–376 (2001)

21. Raschka, S., Mirjalili, V.: Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-learn, and TensorFlow 2. Packt Publishing Ltd. (2019)

22. Dayan, P., Sahani, M., Deback, G.: Unsupervised Learning. The MIT Encyclopedia of The Cognitive Sciences, pp. 857–859 (1999)

23. Wiering, M., van Otterlo, M. (eds.): Reinforcement Learning. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)

24. Kaelbling, L., Littman, M., Moore, A.: Reinforcement learning: a survey. J. Artif. Intell. Res. **4**, 237–285 (1996)

25. Montgomery, D., Peck, E., Vining, G.: Introduction to Linear Regression Analysis. John Wiley & Sons (2021)

26. Kotsiantis, S., Zaharakis, I., Pintelas, P.: Supervised machine learning: a review of classification techniques. Emerg Artifi. Intell. Appl. Comput. Eng. **160**(1), 3–24 (2007)

27. Celebi, M.E., Aydin, K. (eds.): Unsupervised Learning Algorithms. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-24211-8

28. McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. The Bull. Math. Biophys. **5**(4), 115–133 (1943)

29. LeCun, Y., et al.: Handwritten digit recognition with a back-propagation network. In: Touretzky, D. (ed.) Advances in Neural Information Processing Systems, vol. 2. Morgan-Kaufmann (1990)

30. Fukushima, K.: Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern. **36**, 193–202 (1980)

31. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)

32. He, K., Zhang, X., Ren S., Sun J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)

33. Deng, J.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255 (2009)

34. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vision **115**(3), 211–252 (2015)

35. Thontirawong, P., Chinchanachokchai, S.: Teaching artificial intelligence and machine learning in marketing. Mark. Educ. Rev. **31**(2), 58–63 (2021)

36. Miller E., Ceballos H., Engelmann B., Schiffler A., Batres R., Schmitt J.: Industry 4.0 and International Collaborative Online Learning in a Higher Education Course on Machine Learning, Machine Learning-Driven Digital Technologies for Educational Innovation Workshop, pp. 1–8 (2021)

37. Huang L., Ma K.: Introducing machine learning to first-year undergraduate engineering students through an authentic and active learning labware. In: IEEE Frontiers in Education Conference (FIE), pp. 1–4 (2018)

38. Kaggle Homepage. https://www.kaggle.com/c/titanic/data. Last accessed 28 Jan 2021

39. Kaggle Homepage: https://www.kaggle.com/camnugent/california-housing-prices. Last accessed 28 Jan 2021

40. RePDA Homepage: https://app.conagua.gob.mx/consultarepda.aspx. Last accessed 28 Jan 2021

41. Kaggle Homepage: https://www.kaggle.com/hellbuoy/online-retail-k-means-hierarchical-clustering/data. Last accessed 28 Jan 2021

42. Dask Homepage: https://dask.org/. Last accessed 28 Jan 2021

43. Hadoop Homepage: https://hadoop.apache.org/. Last accessed 28 Jan 2021

44. RUOA UNAM Homepage: https://www.ruoa.unam.mx/index.php?page=estaciones. Last accessed 28 Jan 2021

45. Kaggle Homepage: https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes. Last accessed 28 Jan 2021

46. Webmosquito Homepage: http://basurae.iies.unam.mx/webmosquito/html/. Last accessed 28 Jan 2021

47. Kaggle Homepage: https://www.kaggle.com/khoongweihao/covid19-xray-dataset-train-test-sets. Last accessed 28 Jan 2021

48. Kaggle Homepage: https://www.kaggle.com/arpitjain007/game-of-deep-learning-ship-datasets. Last accessed 28 Jan 2021

49. Kaggle Homepage: https://www.kaggle.com/rmisra/news-headlines-dataset-for-sarcasm-detection. Last accessed 28 Jan 2021

50. Kaggle Homepage: https://www.kaggle.com/gpiosenka/sports-classification. Last accessed 28 Jan 2021

51. EscuelaMLDL Homepage: http://www.escuelamldl.enesmorelia.unam.mx/index.php/es/. Last accessed 28 Jan 2022

Acknowledgment

We are grateful for the support of the Instituto de Investigaciones en Ecosistemas y Sustentabilidad (IIES), the CA TRATEC - PRODEP of the Universidad Tecnológica de Morelia (UTM), the TecNM campus Morelia, the Escuela Nacional de Estudios Superiores (ENES), UNAM Campus Morelia, and DGAPA UNAM PAPIME PE106021. Especially thanks to MGTI. Atzimba G. López M., MTI. Alberto Valencia G., Eng. Oscar Álvarez, MTI. Pablo García C. and Eng. Javier Huerta S., for their technical support, comments, and analysis of the statistical calculations. We thank Eng. Alfredo A. Aviña, for his help in applying the survey and Web page support.

Author information

Authors and Affiliations

Escuela Nacional de Estudios Superiores Unidad Morelia, UNAM Campus Morelia, 58190, Morelia, Michoacán, Mexico

Sergio Rogelio Tinoco-Martínez, Froylan Hernández-Rendón, Mariana Michell Flores-Monroy & Bruce Hiram Ginori-Rodríguez

Instituto de Investigaciones en Ecosistemas y Sustentabilidad, UNAM Campus Morelia, 58190, Morelia, Michoacán, Mexico

Heberto Ferreira-Medina

Tecnológico Nacional de México, Campus Morelia, DSC, Morelia, Michoacán, Mexico

Heberto Ferreira-Medina

Departamento de TI, Universidad Tecnológica de Morelia. Cuerpo académico TRATEC-PRODEP. Morelia, 58200, Morelia, Michoacán, México

José Luis Cendejas-Valdez

Corresponding authors

Correspondence to Heberto Ferreira-Medina or José Luis Cendejas-Valdez.

Editor information

Editors and Affiliations

Saga University, Saga, Japan

Kohei Arai

Rights and permissions

Reprints and Permissions

Copyright information

About this paper

Cite this paper

Tinoco-Martínez, S.R., Ferreira-Medina, H., Cendejas-Valdez, J.L., Hernández-Rendón, F., Flores-Monroy, M.M., Ginori-Rodríguez, B.H. (2023). How to Improve the Teaching of Computational Machine Learning Applied to Large-Scale Data Science: The Case of Public Universities in Mexico. In: Arai, K. (eds) Intelligent Systems and Applications. IntelliSys 2022. Lecture Notes in Networks and Systems, vol 544. Springer, Cham. https://doi.org/10.1007/978-3-031-16075-2_1

.RIS .ENW .BIB

DOI

https://doi.org/10.1007/978-3-031-16075-2_1

Published
01 September 2022

Publisher Name
Springer, Cham

Print ISBN
978-3-031-16074-5

Online ISBN
978-3-031-16075-2

eBook Packages
Intelligent Technologies and Robotics
Intelligent Technologies and Robotics (R0)

Heberto Ferreira Medina <hferreira@cieco.unam.mx>

Acceptance Notification : Intelligent Systems Conference (IntelliSys) 2022

IntelliSys Conference <IntelliSys@saiconference.com> 24 de febrero de 2022, 5:27
Para: Heberto Ferreira Medina <hferreira@iies.unam.mx>, stinoco@enesmorelia.unam.mx, luis.cendejas@ut-morelia.edu.mx, fhernandez@enesmorelia.unam.mx, michellmonroy18@gmail.com, bruginrod@gmail.com

Dear Heberto Ferreira Medina, Sergio Rogelio Tinoco-Martínez, José Luis Cendejas-Valdez, Froylán Hernández-Rendón, Mariana Michell Flores-Monroy, Bruce Hiram Ginori-Rodríguez,

Congratulations, your submitted paper "How to Improve the Teaching of Computational Machine Learning Applied to Large-Scale Data Science: The Case of Public Universities in Mexico" has been reviewed and accepted for presentation at the Intelligent Systems Conference (IntelliSys) 2022, to be held from 1-2 September 2022 in Amsterdam, The Netherlands.

We are of course aware that the situation regarding COVID-19 is a cause for apprehension - virtual participation (with reduced registration) is available, for anyone who cannot or chooses not to travel.

Intelligent Systems Conference (IntelliSys) 2022 will focus on areas of intelligent systems and artificial intelligence and how it applies to the real world. IntelliSys provides a leading international forum that brings together researchers and practitioners from diverse fields with the purpose of exploring the fundamental roles, interactions as well as practical impacts of Artificial Intelligence.

In order to attend, present and publish your paper, please register online at https://saiconference.com/IntelliSys2022/Register (Registration closes March 15, 2022). We also accept fees via bank/wire transfer.

IntelliSys 2022 proceedings will be published in the Springer series "Lecture Notes in Networks and Systems" and submitted for consideration to Web of Science, SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH and SCImago.

The Conference Board has decided that the reviewers' feedback and invitation letter for visa applications (if necessary) will be emailed to the author(s) after the registration process. If you would like to receive the reviewer feedback for representation at your university/organization, please feel free to contact us.

Again, congratulations and I look forward to your participation!

Regards,
Kohei Arai
Program Chair
IntelliSys Conference

View IntelliSys 2021 Recap

How to Improve the Teaching of Computational Machine Learning Applied to Large-Scale Data Science: The Case of Public Universities in Mexico

Sergio Rogelio Tinoco-Martínez[1], Heberto Ferreira-Medina[2(*)], José Luis Cendejas-Valdez[3(*)], Froylan Hernández-Rendón[1], Mariana Michell Flores-Monroy[1], Bruce Hiram Ginori-Rodríguez[1]

[1] Escuela Nacional de Estudios Superiores Unidad Morelia, UNAM campus Morelia, Morelia, Michoacán, México, 58190, stinoco@enesmorelia.unam.mx, fhernandez@enesmorelia.unam.mx

[2] Instituto de Investigaciones en Ecosistemas y Sustentabilidad, UNAM campus Morelia, Morelia, Michoacán, México, 58190. Tecnológico Nacional de México, Campus Morelia, DSC. hferreir@iies.unam.mx

[3] Departamento de TI, Universidad Tecnológica de Morelia. Cuerpo académico TRATEC-PRODEP. Morelia, Michoacán, México. 58200, luis.cendejas@ut-morelia.edu.mx

*hferreir@iies.unam.mx,*luis.cendejas@ut-morelia.edu.mx

Abstract. Teaching along with training on Machine Learning (ML) and Big Data in Mexican universities has become a necessity that requires the application of courses, handbooks, and practices that allow improvement in the learning of Data Science (DS) and Artificial Intelligence (AI) subjects. This work shows how the academy and the Information Technology industry use tools to analyze large volumes of data to support decision-making, which is hard to treat and interpret directly. A solution to some large-scale national problems is the inclusion of these subjects in related courses within specialization areas that universities offer. The methodology in this work is as follows: 1) Selection of topics and tools for ML and Big Data teaching, 2) Design of practices with application to real data problems, and 3) Implementation and/or application of these practices in a specialization diploma. Results of a survey applied to academic staff and students are shown. The survey respondents have already taken related courses along with those specific topics that the proposed courses and practices will seek to strengthen, developing needed skills for solving problems where ML/DL and Big Data are an outstanding alternative of solution.

Keywords: Machine Learning, Deep Learning, Big Data, Data Science, Teaching Skills.

1 Introduction

The use of tools that allow the analysis of large volumes of data has allowed exact sciences to play an important role for decision-making in organizations [1]. In the Bachelor of Information Technologies for Sciences (ITCs) of the Escuela Nacional de

Estudios Superiores (ENES) Morelia, Mexico, there are subjects related to Data Science (DS) [2] that are included in the curriculum starting from the 6th semester, known as subjects of the deepening area, and that represent a challenge for students when trying to put the theory learned into practice, in addition to lacking the necessary tools for its application in real problems. The need for teachers and students to know new frontiers in Artificial Intelligence (AI) is observed, specifically in the application of mathematical models of Machine Learning (ML). ML is the branch of AI that is responsible for developing techniques, algorithms, and programs that give computers the ability to learn. A machine learns each time it changes its structure, programs, or data, based on input or in response to external information, in such a way that better performance is expected in the future [3].

In [4] Deep learning (DL) is used to explain new architectures of Neural Networks (NN) that are capable of learning. DL is a class of ML techniques that exploit many layers of nonlinear processing for extraction and transformation of supervised and unsupervised features and pattern analysis and classification [5]. The 21st century has become the golden age for AI; this is due, in large part, to a greater computing capacity and the use of GPUs to speed up the training of these systems, with the ingestion of large amounts of data. Currently, numerous frameworks have ML and DL tools implemented, such as PyTorch [6], fast.ai [7], TensorFlow [8], Keras [9], DL4J [10], among others. Some of the main uses of DL today are, for example, identifying brand names and company logos in photos posted on social media, real-time monitoring of reactions on online channels during product launches, ad recommendation and prediction of preferences, as well as identification and monitoring of customer confidence levels, among others. The use of AI has allowed a better understanding of genetic diseases and therapies, the analysis of medical images (such as X-rays and magnetic resonance imaging) increasing the accuracy of diagnosis in less time and at a lower cost than traditional methods [11]. The DL forms a subcategory of the ML. To differentiate it from the rest of ML algorithms, it uses the fact that large-scale NNs allow a machine to learn to recognize complex patterns by itself, which is difficult to achieve with them [12]. A NN is made up of several layers or levels and a certain number of neurons in each of them, which constitute the processing unit, whose mathematical model allows having several data inputs and an output that is the weighting of their inputs [13] [14]. The connections of several neurons within an NN constitute a powerful parallel computation tool, capable of delivering approximate and non-definitive outputs. Furthermore, NNs can be structured in various ways and can be trained with various types of algorithms [15]. On the other hand, the Internet of Things (IoT) and industry 4.0 have required the introduction of autonomous and intelligent machinery in the industrial sector [16]. In this industry, Convolutional Neural Networks (CNN) are applied, which are a type of DL that is inspired by the functioning of the visual cortex of the human brain and differs from the other NNs by the fact that each of the neurons of the layers that compose it does not receive incoming connections from all the neurons of the previous layer, but only from some of them. This simplifies the learning of the network, generating lower computational and storage costs. All of the above mentioned makes DL models more accurate [12].

The main contribution of this paper is to present a proposal to improve the learning of ML, DL, and Big Data subjects with practical training focused on real-life use cases. The rest of the paper is organized as follows: Section 2 describes the work related to

AI (ML, DL, and Big Data) and the implementation of its teaching in courses or diplomas oriented to data science. It shows the difficulty of the students in learning and applying the topics of ML and DL in data analysis. This problem must be solved with a practical approach or orientation. Section 3 describes the methodology used to develop the improvement proposal. Section 4 presents the results, as well as a brief discussion of them. Finally, Section 5 presents the conclusions obtained from the developed proposal.

2 Related Work

In current terms [17] explains that within AI there is a branch called ML whose purpose is to improve the performance of algorithms through their experience. The ML uses statistics, computer science, mathematics, and computational sciences having its foundation in data analysis. The definition coincides with other proposals that consider ML as the technique of creating systems that are capable of learning by themselves, using large volumes of data, making them suitable for analysis and, thus, being able to predict future behavior [18]. Regarding the ML as an area [12] points out that there are different approaches for the design of these systems. The approaches are divided into supervised, unsupervised and by reinforcement, if the system is trained under human supervision or not (the most used division); online or offline learning, whether the system can learn on the fly or not; and, finally, in instance-based learning or model-based learning, if the system detects training patterns or if it compares new data against existing data.

The supervised ML approach is used when you already have data, and you know the response you want to predict. This knowledge is then used to predict the labels of new data whose label is unknown. The main problems that are solved with this type of learning are regression (predicting the future value of a given element, whose values can only be numerical, from relevant characteristics and previous values [19]) and classification (assigning a label to a given element, from a discrete set of possibilities [20]).

In the unsupervised ML approach, the data is not labeled, this means that the system must learn by itself without being told if the classification is correct or not [21]. To classify the data, a grouping technique is used whose objective is to combine data whose characteristics are like each other [22].

Regarding the reinforcement ML approach, the goal is that the system learns in an environment in which the only feedback consists of a scalar reward, which can be positive or negative (punishment) [23]. That is, the system receives in each iteration a reward and the current state of the environment, then takes an action according to these inputs and what results is considered as an output, which will change the state in the next iteration [24].

The arguably most used algorithms of supervised ML are linear or logistic regression, for regression [25]; and decision trees, k-nearest neighbors, support vector machines or artificial NNs, for classification [26]. The algorithms for unsupervised ML are k-means, visualization and dimensionality reduction or association rules [27].

In relation to artificial NNs, they serve to solve both regression and classification problems and even some unsupervised learning problems. Due to its versatility and performance that have recently surpassed even human performance (at the cost of having example data in large quantities that the IoT has allowed to obtain).

The study of DL began in 1943 as a computer model inspired by the NNs of the human brain [28], however, it was not until 1985 that in [29] it was demonstrated that backpropagation in a NN could provide distribution representations of great utility for learning, generating a reborn interest in the area.

In [30] the first practical demonstration of backpropagation is provided. The team combined CNNs with backpropagation to read handwritten digits in a system that, finally, was used to read handwritten check numbers. The model was based on the hierarchical multi-layer design (CNNs) inspired by the human visual cortex of the Neocognitron, introduced in 1979 [31].

In the late 1990s, the problem of the fading or exploding gradient was detected in DL models. The problem originates in the activation functions of artificial neurons whose gradient (based on the derivative) decreased when calculated in each layer, until it reached practically zero (or tended to an infinite value); which implies loss of learning. The proposed solution is to store the gradient within the network itself [32] or to make it enter a layer and simultaneously avoid it through skip connections [33].

In [32] Recurrent NNs (RNNs) of the Long Short-Term Memory (LSTM) type are proposed, which specialize in the analysis and prediction of time series and problems that must recall previous states, such as Natural Language Processing (NLP). In addition, this architecture allows to solve the gradient problems mentioned above.

In 2009 ImageNet [34] was launched, a free, tagged database of more than 14 million images, which features a thousand different categories of objects as varied as 120 dog breeds. With this resource and coupled with the computing power that the evolution of GPUs already had by 2011, it became possible to train CNNs without the previous training (layer by layer) and considering architectures with an increasing number of these (hence the term DL).

The examples that can be mentioned of the efficiency and speed that DL algorithms have achieved are the computer vision algorithms that were winners in the ImageNet Large Scale Visual Recognition Challenge (LSVRC) [35] between the years of 2012 up to 2017 (all CNNs architectures), whose main challenge is based on the classification of images from the ImageNet database and that, as of 2017, it is considered solved in practice and with superior performance to that of the human being.

To our knowledge, there are very few works in the recent literature related to the improvement of AI teaching. Some of the most representative ones are mentioned below, as a review of the approaches they address and that are aimed at teaching AI as a secondary objective or use case.

According to [36] ML is a discipline that focuses on building a computer system that can improve itself using experience. ML models can be used to detect patterns from data and recommend strategic marketing actions, showing how educators can improve the teaching of these topics using the AI approach. The availability of Big Data has created opportunities and challenges for professionals and academics in the area. Therefore, study programs must be constantly updated to prepare graduates for rapidly changing trends and new approaches. In [37] it is described that the pandemic caused by the COVID-19 virus, the advent of Industry 4.0 confronts graduate students with the need to develop competencies in ML, which are applied to solve many industrial problems that require prediction and classification, and the availability and management of large amounts of data. The proposal of how to apply AI practices in a virtual laboratory is shown, in addition to evaluating the performance of students in this type of environment. In the same way, in [38], an innovative practice of teaching applied ML to first-year multidisciplinary engineering university students is proposed, using a learning tool that consists of a public repository in the cloud and a course project. A set of practices for ML and how to apply it in real cases is offered as a use case for online collaborative work. The inclusion of DL and Big Data is mentioned as future work.

In Mexico, there are many academic degrees oriented to DS, ML, and DL. However, there are no uniform curricula on the areas of knowledge, topics, and tools that students require. In this paper, we offer an alternative way to solve this problem based on the experience of a public university such as the ENES Morelia - UNAM.

3 Methodology

Based on the review of the literature, a series of steps were generated, which allowed us to obtain the level of knowledge that the ENES Morelia population has about ML/DL and thus be able to define axes that support the design of the pedagogical strategy that will give rise to the proposal of a uniform curriculum through courses of practical experience. The present research is characterized by being a study of type: 1) exploratory, 2) descriptive, 3) correlational and 4) pre-experimental to have a case study through a single measurement. To this end, a survey was generated that was applied to a population made up of professors, students, and researchers of the UNAM (Morelia, Michoacan campus). The methodology followed for this work is shown in Fig. 1. Results of the analysis of the application of this survey and the monitoring of test groups are described in the following sections.

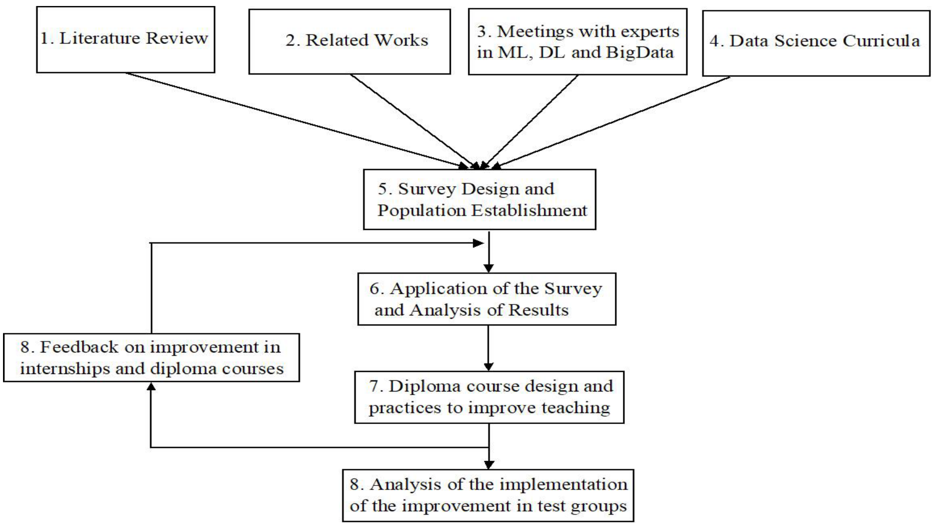


Fig. 1. Block diagram of the methodology used in this work.

3.1 Population

Primarily the survey was created through the "e-Encuesta®" Web platform and distributed via email and social media to randomly selected individuals affiliated with the campus mentioned before. In the first place, the sample was calculated using the finite population method based on 600 people. This sample has a confidence interval of 95% and a margin of error of 10%, as shown in Table 1. The link to the survey within the Web platform was distributed through the official email of students and teachers. It was validated that the information in each of the answers was consistent and complete.

Table 1. Population sample.

Description	Value
Population size	600
Trust level	95%
Margin of error	10%
Sample size	83

3.2 Survey

An eight-question survey was generated from a critical review of the literature related to DS of ML, DL and Big Data. Experts on the subject validated the selected questions. Its measurement was: a) different options, b) dichotomous responses, and c) Likert scale. The survey was refined by dividing it into four axes: Axis I: Machine Learning; Axis II: Deep Learning; Axis III: Big Data; and Axis IV. Tools, as shown in Table 2. Likert scale applied was: Very important, Important, Neutral, Less important, Nothing important, I do not know.

Table 2. Survey questions and format.

#	Question description	Axis	Type
Q1	UNAM account or employee number, your name if you do not have them	-	Options
Q2	Bachelor's degree you are studying; 2.1 Sciences, 2.2 Agroforestry, 2.3 Environmental Sciences, 2.4 Sustainable Materials Science, 2.5 Ecology, 2.6 Social Studies and Local Management, 2.7 Geosciences, 2.8 Geohistory, 2.9 Information Technologies in Sciences, 2.10 Other	-	Options
Q3	Semester you are studying (1-12), does not apply to teachers and researchers	-	Number

#	Question description	Axis	Type
Q4	You consider the following topics related to ML to be: 4.1 Data Science, 4.2 Web Scraping, 4.3 Data Wrangling, 4.4 Machine Learning, 4.5 Data Mining, 4.6 Ensemble Learning, 4.7 Data visualization, 4.8 ML: supervised/unsupervised, 4.9 Binary and multiclass classification, 4.10 EDA, 4.11 Clustering, 4.12 ML model, 4.13 ML evaluation: underfitting, overfitting, 4.14 Cross validation, 4.15 Hyperparameters, regularization, feature engineering, 4.16 PCA	I	Likert options
Q5	You consider the following topics related to DL to be: 5.1 NN Shallow & Deep, 5.3 CNN, 5.3 RNN, 5.4 Transfer Learning & Fine-Tuning, 5.5 Dropout, 5.6 Data Augmentation, 5.7 Batch Normalization	II	Likert Options
Q6	You consider the following topics related to Big Data to be: 6.1 Concept, 6.2 Model Scaling, 6.3 Large-Scale Analytics, 6.4 Distributed File System, 6.5 Map-Reduce	III	Likert Options
Q7	Skills you have in handling the following tools is: 7.1 TensorFlow, 7.2 Spark, 7.3 Keras, 7.4 Fast.ai, 7.5 PyTorch, 7.6 HDFS, 7.7 Kafka, 7.8 Python, 7.9 Scikit-Learn	IV	Likert Options
Q8	Is it important to include some additional topics related to ML, DL and Big Data, not mentioned above?	-	Open

3.3 Data Analysis

In the second place, with the information obtained from the survey, descriptive analyses were generated, where the reliability study was carried out applying Cronbach's Alpha obtaining as a result 0.956 and demonstrating that the information obtained is consistent. Third, the study of correlations was applied using Pearson's bivariate and selecting only the correlations obtained at the high and very high levels [0.7-0.93], shown in Fig. 2.

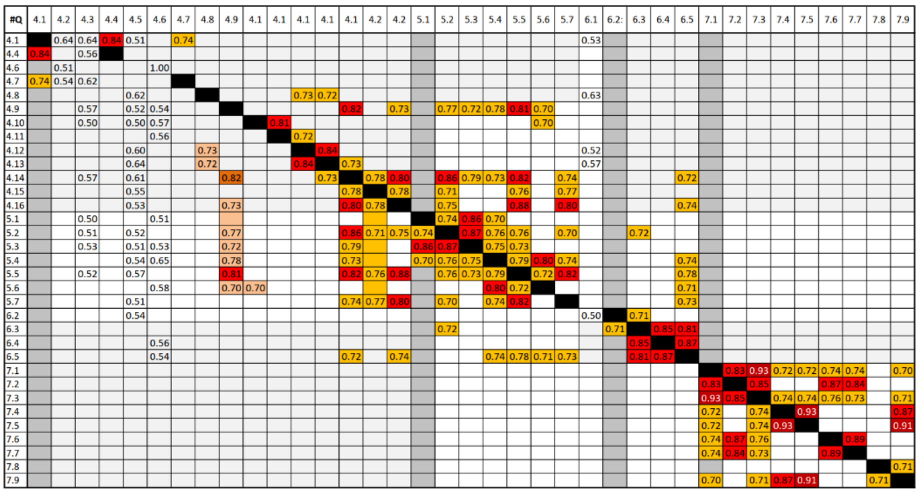
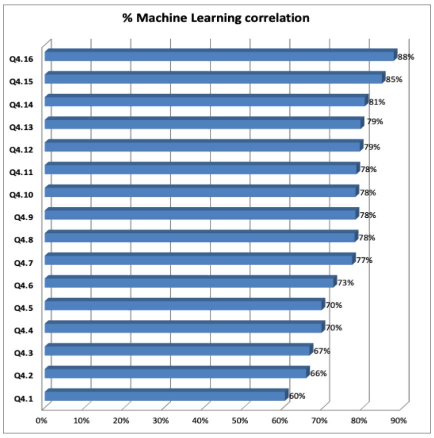
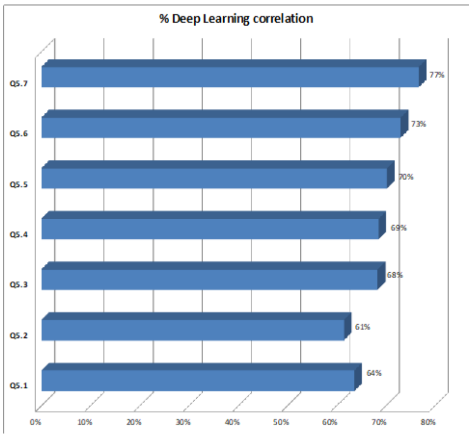


Fig. 2. Correlation matrix (heat map). See Table 1, for tag details.

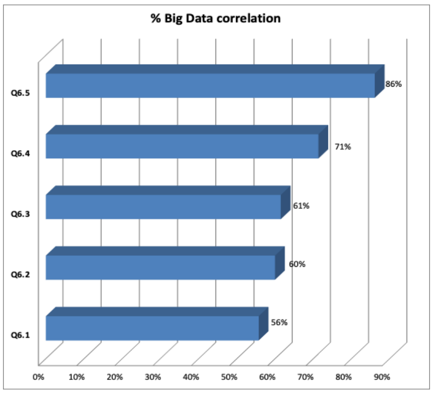
According to the analysis of correlations, we identify areas of opportunity according to percentage of importance (scale) that the respondents answered. In Fig. 3 this importance is shown.



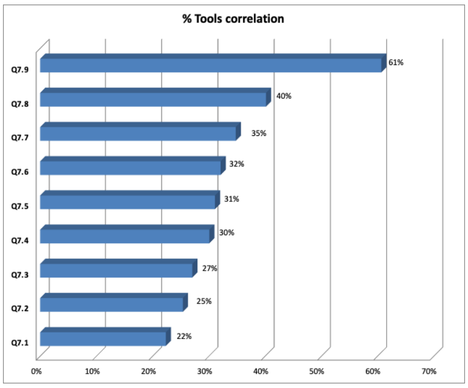
a) ML importance level



b) DL importance level



c) Big Data importance level



d) Tools importance level

Fig. 3. Level of importance (Likert) according to survey respondents; a) ML, b) DL, c) Big Data and d) Tools. See Table 1, for tag details.

4 Results and Discussion

According to the developed survey, a certain lack of knowledge of the respondents was observed in some topics. In Fig. 4 topics are shown by axis, ordered by level of unfamiliarity: I do not know (dnK), Nothing important (NImp), Less Important (LImp), Neutral, Important (Imp), Very Important (VImp). And, for tools, scale is: I do not know (dnK), Short, Half, and High.

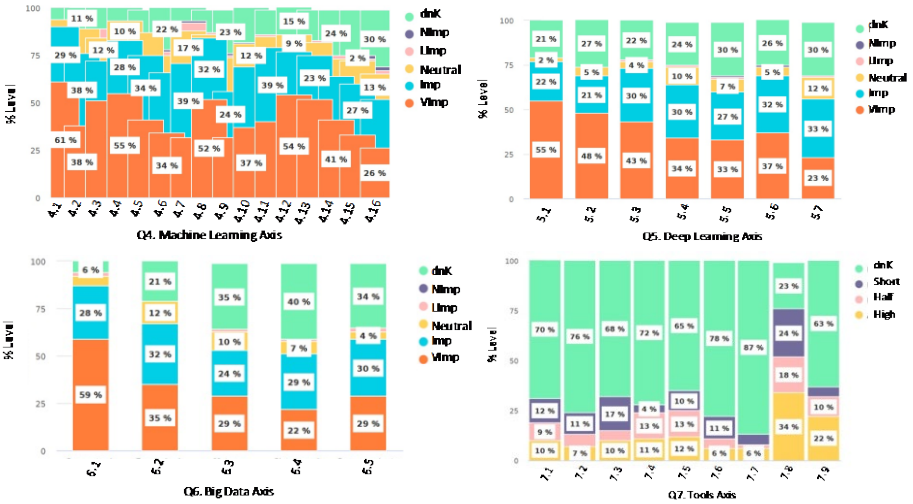


Fig. 4. Levels of unfamiliarity by axes. See Table 1, for tag details.

Based on the analysis of these results and considering the classic progression in recent literature, regarding the teaching of basic topics of ML, DL, and Big Data; coupled with our personal experience in teaching courses on these topics, at the undergraduate level and more of them aimed at teachers in the area of Information Technology, it is proposed to improve learning with practical training to strengthen those topics with the greatest lack of knowledge (dnK level of unfamiliarity in Fig. 4). This practical knowledge is shown in Tables 3 and 4 as a series of practices we recommended to take advantage of these areas for improvement. It was observed that respondents prefer an intervention oriented towards the practical application of knowledge.

Table 3. Proposed ML practices.

#	Name	Dataset	Evaluation Metric	Description
1	Classification using decision trees	Titanic passengers [39]	Accuracy and/or Fbeta Metrics	Build a decision tree for the survival analysis of Titanic passengers (Classification)
2	Housing cost prediction	California Housing [40]	RMSE and/or MAE	Build a real estate cost prediction model. (Linear Regression/Logistic Regression)
3	k-Nearest Neighbors	Water wells [41]	Precision Score Fbeta	Build a prediction model of water well uses. (Supervised).
4	k-Means	Online Retail K-means & Hierarchical Clustering [42]	Not applicable	Design a model to classify the transactions of a bank's customers. (No supervised).

#	Name	Dataset	Evaluation Metric	Description
5	Installing and using Dask [43]	Not applicable	Not applicable	Show Dask installation and how it is used for Big Data manipulation
6	Installing and Using HDFS [44]	Not applicable	Not applicable	Teaching how the installation of HDFS and its basic use is carried out.
7	Weather forecasting	RUOA (UNAM, 2015) [45]	RMSE and/or MAE	Analyze climate data from the RUOA to predict weather on a daily horizon. (Linear Regression)
9	Car Price Prediction	100,000 UK Used Car Data set [46]	RMSE and/or MAE	Analyze car data to estimate prices. (Multiple regression)
10	Special case	Public data information	Several	Analyze data to apply the best strategy to solve a problem

Table 4. Proposed DL practices.

#	Name	Dataset	Evaluation Metric	Description
1	Binary classification with CNN	800 images of mosquitoes, UNAM [47]	Accuracy	Differentiate between species Aedes *Albopictu*s and Aedes *Aegypti*. (Visualization)
2	Binary classification with CNN	Covid-19 Pneumonia Screening [48]	Precision & Confusion Matrix.	X-ray tomography analysis for identification of lungs affected by the SARS-CoV-2 virus.
3	CNN & Data Augmentation	Ship Classification [49]	Precision & Confusion Matrix.	Classification of 6,252 images of ships (5 categories)
4	RNN	Sarcasm Detection [50]	Accuracy and precision.	Identify news titles that are sarcastic or satirical. (NLP)
5	Installing and using PyTorch over Dask	Not applicable	Not applicable	Installation and use of PyTorch in Dask
6	Transfer Learning	Sports Images [51]	Precision and accuracy	Classification of sports images.

5 Conclusions

By the experience of practical teaching to a mix of students and teachers of the Morelia campus of the UNAM university, divided into two heterogeneous groups, concerning applying the proposed practices, two courses were offered according to the diploma described below [52]:

MODULE I. Machine Learning (ML). "Theory and Practice for the Improvement of the Teaching of ML Applied to Data Science". Topics: 1. Artificial Intelligence and Machine Learning, 2. Phases of an ML Project, 3. Regression Methods, 4. Classification Methods, 5. Prediction Methods, 6. Supervised Learning, 7. Unsupervised Learning, 8. Metrics. Practices to be Developed: See Table 3.

MODULE II. Deep learning (DL). "Theory and Practice for the Improvement of the Teaching of DL Applied to Data Science". Topics: 1. Artificial Intelligence and Deep Learning (DL), 2. Phases of a DL Project, 3. Convolutional Neural Networks (CNNs), 4. Learning Transfer, 5. Recurrent Neural Networks (RNNs), 6. Visualization and Treatment of Natural Language Processing (NLP). Practices: See Table 4.

Tools used in the diploma: Anaconda Python, Scikit-Learn, Matplotlib, Dask, PyTorch, fast.ai 2, TensorFlow, Keras, HDFS, among others.

At the end of the first course where the intervention was carried out, it was observed that 50% of the attendees, of a total of 40, had various problems solving practices. These problems are shown as percentages of solved practices in Fig. 5. In this figure, the expected results (according to our experiences in previous courses) are compared against the actual results, as practices solved and delivered by the attendees. In addition, the figure shows the efficiency of teaching according to:

$$\%efficiency = 100 * (solved\ practices - expected\ practices)/expected\ practices \quad (1)$$

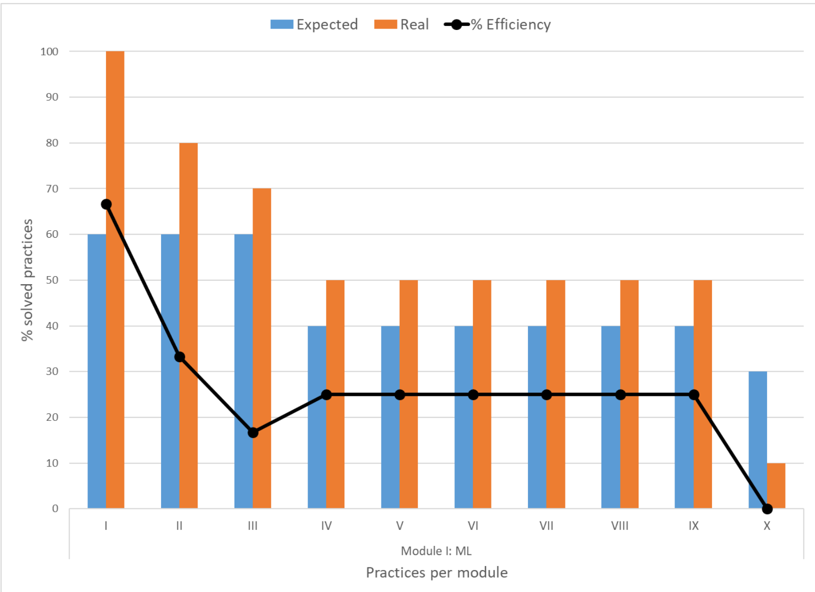


Fig. 5. ML teaching effectiveness.

We observed that students quit working on the most complex practices of ML. Reasons were the increase in data analysis tasks, on top of having to apply statistical and mathematical theory using a programming language (Python). The solution to these problems is to give higher priority to practice with real data than to abstract theory.

In works similar to this one, the teaching of ML is used only as a use case to address education in other topics in real cases; indeed, with the AI approach, but no improvements are made to the teaching of ML itself, nor experiences on how to improve the teaching of data science are included. The practical proposal in this work allows for establishing a more complete and broad curriculum, concerning the fact that it includes not only the ML but the DL, the Big Data, and the computer tools associated with data science as well.

Our proposal is still under development and, among other issues, it is necessary to evaluate the efficiency of teaching according to this practical approach in a DL course, as well as including other tools that may help facilitate the learning of data science. All this, in addition to including learning platforms as well as other tools that can help facilitate data science learning, such as collaborative learning platforms in the cloud.

6 Acknowledgment

We are grateful for the support of the Instituto de Investigaciones en Ecosistemas y Sustentabilidad (IIES), the CA TRATEC - PRODEP of the Universidad Tecnológica de Morelia (UTM), the TecNM campus Morelia, the Escuela Nacional de Estudios Superiores (ENES), UNAM Campus Morelia, and DGAPA UNAM PAPIME PE106021. Especially thanks to MGTI. Atzimba G. López M., MTI. Alberto Valencia G., Eng. Oscar Álvarez, MTI. Pablo García C. and Eng. Javier Huerta S., for their technical support, comments, and analysis of the statistical calculations. We thank Eng. Alfredo A. Aviña, for his help in applying the survey and Web page support.

References

1. Haenlein, M.; Kaplan, A.: A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. California Management Review, 61 (4), pp. 5–14. (2019)
2. ENES-UNAM Homepage, http://www.enesmorelia.unam.mx/, last accessed 2021/01/18.
3. Nilsson, N. J.: Introduction to Machine Learning. Not published, Stanford, CA. (1996).
4. Rumelhart, D., Hinton, G., Williams, R. Learning representations by back-propagating errors. Nature, vol. 323, pp. 533–536 (1986).
5. Deng, L., Yu, D.: Deep Learning: Methods and Applications, Foundations and Trends in Signal Processing: vol. 7, pp. 197-387 (2014).
6. PyTorch Homepage, https://pytorch.org/, last accessed 2021/01/28.
7. Fast.ai Homepage, https://www.fast.ai/, last accessed 2021/01/28.
8. TensorFlow Homepage, https://www.tensorflow.org/, last accessed 2021/01/28.
9. Keras Homepage, https://keras.io/, last accessed 2021/01/28.
10. Deeplearning4j Homepage, https://deeplearning4j.konduit.ai, last accessed 2021/01/28.

11. El Naqa, I., Murphy, M. J.: What is Machine Learning? In Machine Learning in Radiation Oncology (pp. 3–11). Springer. (2015).
12. Géron, A.: Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media. (2019).
13. Hagerty, J., Stanley R., Stoecker W.: Medical image processing in the age of deep learning Proceedings of the 12th international joint conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP), pp. 306-311. (2017).
14. Basheer, I., Hajmeer M.: Artificial neural networks: fundamentals, computing, design, and application, Journal of Microbiological Methods, Volume 43, Issue 1, pp. 3-31, ISSN 0167-7012 (2000).
15. Lévy J., Flórez R., Rodríguez J.: Las redes neuronales artificiales. Netbiblo: Tirant lo Blanch. 1a. Edición. (2008).
16. Lasi, H., Fettke, P., Feld, T., Hoffmann, M.: Industry 4.0. Business & Information Systems Engineering, 6 (4), 239-242. (2014).
17. Jordan, M., Mitchell, T.: Machine Learning: Trends, Perspectives, and Prospects. Science, 349 (6245), pp. 255–260. (2015).
18. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of Machine Learning. MIT Press. (2018).
19. Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., Li, B.: Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. En 2018 IEEE Symposium on Security and Privacy (SP) (pp. 19–35). (2018).
20. Li, W., Han, J., Pei, J.: CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules. En Proceedings 2001 IEEE. International Conference on Data Mining (pp. 369–376). (2001).
21. Raschka, S., Mirjalili, V.: Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-learn, and TensorFlow 2. Packt Publishing Ltd. (2019).
22. Dayan, P., Sahani, M., Deback, G.: Unsupervised Learning. The MIT Encyclopedia of The Cognitive Sciences, pp. 857–859. (1999).
23. Wiering, M., Van, M.; Reinforcement Learning. Adaptation, Learning, and Optimization, 12 (3). (2012).
24. Kaelbling, L., Littman, M., Moore, A.: Reinforcement Learning: A Survey. Journal of Arti cial Intelligence Research, 4, pp. 237–285. (1996).
25. Montgomery, D., Peck, E., Vining, G.: Introduction to Linear Regression Analysis. John Wiley & Sons. (2021).
26. Kotsiantis, S., Zaharakis, I., Pintelas, P.: Supervised Machine Learning: A Review of Classi cation Techniques. Emerging Articial Intelligence Applications in Computer Engineering, 160 (1), pp. 3–24. (2007).
27. Celebi, M., Aydin, K.: Unsupervised Learning Algorithms. Springer. (2016).
28. McCulloch, W., Pitts, W.: A logical Calculus of the Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, pp. 115–133 (1943).
29. Rumelhart, D., Hinton, G., Williams, R.: Learning representations by back-propagating errors. Nature, 323:533–536 (1986).
30. LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard W., Jackel, L.: Handwritten digit recognition with a back-propagation network. In D. Touretzky, editor, Advances in Neural Information Processing Systems, volume 2. Morgan-Kaufmann. (1990).
31. Fukushima, K.: Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. Biological Cybernetics, 36:193–202 (1980).

32. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Comput; 9 (8): 1735–1780 (1997).
33. He, K., Zhang, X., Ren S., Sun J.: Deep Residual Learning for Image Recognition, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778 (2016).
34. Deng, J.: ImageNet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. (2009).
35. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S, Ma, S., Huang, Z. Karpathy, A., Khosla, A., Bernstein, M., Berg, A., Fei-Fei L.: ImageNet Large Scale Visual Recognition Challenge. IJCV, 2015.
36. Thontirawong, P., Chinchanachokchai, S.: Teaching Artificial Intelligence and Machine Learning in Marketing, Marketing Education Review 31(2), pp. 58-63 (2021).
37. Miller E., Ceballos H., Engelmann B., Schiffler A., Batres R., Schmitt J.: Industry 4.0 and International Collaborative Online Learning in a Higher Education Course on Machine Learning, Machine Learning-Driven Digital Technologies for Educational Innovation Workshop, pp. 1-8 (2021).
38. Huang L., Ma K.: Introducing Machine Learning to First-year Undergraduate Engineering Students Through an Authentic and Active Learning Labware, IEEE Frontiers in Education Conference (FIE), pp. 1-4 (2018).
39. Kaggle Homepage, https://www.kaggle.com/c/titanic/data, last accessed 2021/01/28.
40. Kaggle Homepage, https://www.kaggle.com/camnugent/california-housing-prices, last accessed 2021/01/28.
41. RePDA Homepage, https://app.conagua.gob.mx/consultarepda.aspx, last accessed 2021/01/28.
42. Kaggle Homepage, https://www.kaggle.com/hellbuoy/online-retail-k-means-hierarchical-clustering/data, last accessed 2021/01/28.
43. Dask Homepage, https://dask.org/, last accessed 2021/01/28.
44. Hadoop Homepage, https://hadoop.apache.org/, last accessed 2021/01/28.
45. RUOA UNAM Homepage, https://www.ruoa.unam.mx/index.php?page=estaciones, last accessed 2021/01/28.
46. Kaggle Homepage, https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes, last accessed 2021/01/28.
47. Webmosquito Homepage, http://basurae.iies.unam.mx/webmosquito/html/, last accessed 2021/01/28.
48. Kaggle Homepage, https://www.kaggle.com/khoongweihao/covid19-xray-dataset-train-test-sets, last accessed 2021/01/28.
49. Kaggle Homepage, https://www.kaggle.com/arpitjain007/game-of-deep-learning-ship-datasets, last accessed 2021/01/28.
50. Kaggle Homepage, https://www.kaggle.com/rmisra/news-headlines-dataset-for-sarcasm-detection, last accessed 2021/01/28.
51. Kaggle Homepage, https://www.kaggle.com/gpiosenka/sports-classification, last accessed 2021/01/28.
52. EscuelaMLDL Homepage. http://www.escuelamldl.enesmorelia.unam.mx/index.php/es/, last accessed 2022/02/28.