Lecture Notes in Networks and Systems 1364

Vikrant Bhateja
Vazeerudeen Abdul Hameed
Siba K. Udgata
Jinshan Tang *Editors*

Innovations in ICT: Sustainability for Societal and Industrial Impact

Proceedings of 5th International Conference on Data Engineering and Communication Technology (ICDECT 2024), Volume 3

Springer

Lecture Notes in Networks and Systems

Volume 1364

Series Editor

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Advisory Editors

Fernando Gomide, Department of Computer Engineering and Automation—DCA, School of Electrical and Computer Engineering—FEEC, University of Campinas—UNICAMP, São Paulo, Brazil

Okyay Kaynak, Department of Electrical and Electronic Engineering, Bogazici University, Istanbul, Türkiye

Derong Liu, Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, USA
Institute of Automation, Chinese Academy of Sciences, Beijing, China

Witold Pedrycz, Department of Electrical and Computer Engineering, University of Alberta, Alberta, Canada
Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Marios M. Polycarpou, Department of Electrical and Computer Engineering, KIOS Research Center for Intelligent Systems and Networks, University of Cyprus, Nicosia, Cyprus

Imre J. Rudas, Óbuda University, Budapest, Hungary

Jun Wang, Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

The series "Lecture Notes in Networks and Systems" publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, EI Compendex, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

Vikrant Bhateja · Vazeerudeen Abdul Hameed ·
Siba K. Udgata · Jinshan Tang
Editors

Innovations in ICT: Sustainability for Societal and Industrial Impact

Proceedings of 5th International Conference on Data Engineering and Communication Technology (ICDECT 2024), Volume 3

Springer

Editors
Vikrant Bhateja
Department of Electronics Engineering
Faculty of Engineering and Technology (UNSIET)
Veer Bahadur Singh Purvanchal University
Jaunpur, Uttar Pradesh, India

Siba K. Udgata
School of Computer and Information Sciences
University of Hyderabad
Hyderabad, Telangana, India

Vazeerudeen Abdul Hameed
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumpur, Malaysia

Jinshan Tang
Department of Health Administration and Policy, College of Public Health
George Mason University
Fairfax, VA, USA

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-981-96-5219-8 ISBN 978-981-96-5220-4 (eBook)
https://doi.org/10.1007/978-981-96-5220-4

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

If disposing of this product, please recycle the paper.

Preface

The 5th International Conference on Data Engineering and Communication Technology (ICDECT 2024) follows the immense success of its past four editions in various countries. The conference is organized by Asia Pacific University (APU) during September 28–29, 2024, in a hybrid format, allowing both in-person and virtual participation, making it accessible to a worldwide audience.

With the theme "Innovation Paradigms: Knowledge, Intelligence, and Sustainability for Societal, Environmental, and Industrial Impact", the event brought together researchers, academics, and professionals to discuss the latest advancements in technology and their practical applications across industries. The registered papers in this conference have been organized into four proceedings volumes, further contributing to the growing knowledge in data engineering and communication technology. This book, *Innovations in ICT: Sustainability for Societal and Industrial Impact*, is a compilation of papers which brings together cutting-edge research in Information and Communication Technology (ICT) presented at ICDECT 2024.

ICDECT 2024 began with an engaging inaugural session, graced by distinguished leaders including **Prof. Dr. Ho Chin Kuan**, General Chair and Vice Chancellor of Asia Pacific University (APU), **Prof. Ts. Dr. Murali Raman**, Co-chair and Deputy Vice Chancellor of APU, along with Steering Committee Chairs **Prof. Suresh Chandra Satapathy** (KIIT Deemed to be University, India) and **Prof. Siba K. Udgata** (University of Hyderabad, India). Their insightful addresses set the tone for the conference, emphasizing the importance of interdisciplinary research in the ever-evolving fields of data engineering and communication technologies. The first day of the conference was held in a physical mode and was packed with a series of keynote lectures and paper presentations by authors. These provided attendees with an opportunity to explore cutting-edge research in the field. The program also included valuable networking opportunities for the participants with industry experts and other elites in the field of data engineering.

ICDECT 2024 was an excellent platform for new ideas, networking, and collaborative efforts to address the pressing challenges of the modern world. The ICDECT 2024 is honored to host a distinguished lineup of keynote speakers, each offering

deep insights into their fields. Prof. Ishak Bin Aris from Universiti Putra Malaysia, Malaysia, will deliver the opening keynote, titled "The Past, Present and Future Trends of Vehicle Communication System", exploring the evolution and future directions of vehicular communication technologies. Following this, Prof. Angela Lee Siew Hoong from Sunway University, Malaysia, will present "Revolutionizing Healthcare: The Transformative Power of AI", focusing on how artificial intelligence reshapes the healthcare industry and its potential impact. The third keynote speaker, Mr. Aninda Bose, Executive Editor at Springer Nature, London, UK, will share his expertise on "Nuances in Scientific Publishing", discussing the intricate processes behind scientific publications and the evolving landscape of academic publishing. On the second day, Assoc. Prof. Dr. Selvakumar Samuel from Asia Pacific University, Malaysia, will speak on "The Limits of Automated Intelligence", addressing AI's current boundaries and potential future capabilities. Each keynote session promises to offer invaluable perspectives for researchers and professionals in the field.

ICDECT 2024 was a resounding success, thanks to the meticulous efforts and dedication of the organizing committee. We extend our heartfelt congratulations to the organizers for their outstanding work in making this event memorable and impactful. A special thank you to all the authors and attendees, whose active participation and enthusiasm made the conference a remarkable experience. We would also like to express our deepest gratitude to the technical session chairs, whose expertise and guidance played a pivotal role in shaping the technical program. Our appreciation extends to the special session chairs and reviewers, whose tremendous knowledge and diligent efforts in evaluating manuscripts significantly contributed to the quality of the conference.

Jaunpur, India — Vikrant Bhateja
Kuala Lumpur, Malaysia — Vazeerudeen Abdul Hameed
Hyderabad, India — Siba K. Udgata
Virginia, USA — Jinshan Tang

Interpretive Clustering in Vocational Orientation as a Tool for Career Planning

José L. Cendejas Valdez, Gustavo Adolfo López Saldaña, Sergio Rogelio Tinoco Martínez, Heberto Ferreira Medina, Gustavo A. Vanegas Contreras, Jesús L. Soto Sumuano, Juan J. Ocampo Hidalgo, and Yadira Guillén Sánchez

Abstract Choosing a career path is a crucial decision in a high school student's life. With the proliferation of educational options, the process can be overwhelming. This is where machine learning (ML) comes in and specifically, clustering, which is a technique used to group similar data. This approach helps you find a career that fits your interests and skills. Clustering applied in career counseling involves analyzing data related to students' interests, skills, personality, and preferences. ML algorithms

J. L. C. Valdez (✉) · G. A. V. Contreras
CA—TRATEC PRODEP, Master's Degree in Applied Engineering in Technological Innovation, Information Technology Area, Universidad Tecnológica de Morelia, Morelia, México
e-mail: luis.cendejas@ut-morelia.edu.mx

G. A. L. Saldaña
Masters Degree in Applied Engineering in Technological Innovation, Renewable Energy Area, Universidad Tecnológica de Morelia, Morelia, México
e-mail: gustavo.lopez@ut-morelia.edu.mx

S. R. T. Martínez
Information Technologies in Science, Escuela Nacional de Estudios Superiores Unidad Morelia, Universidad Nacional Autónoma de México, Morelia, México
e-mail: stinoco@enesmorelia.unam.mx

H. F. Medina
Institute for Research in Ecosystems and Sustainability, National Autonomous University of Mexico, UNAM Campus Morelia, Morelia, Mexico
e-mail: hferreir@iies.unam.mx

J. L. S. Sumuano
Departamento de TI, CUCEA—Universidad de Guadalajara, Cuerpo académico de TI—PRODEP, Zapopan, Jalisco, Mexico

J. J. O. Hidalgo
Electronics Department of the Basic Sciences and Engineering Division of the Azcapotzalco Unit of the Universidad Autónoma Metropolitana (UAM), Morelia, Mexico
e-mail: jjoh@azc.uam.mx

Y. G. Sánchez
Melchor Ocampo federal cooperative high school, Teaching—educational guidance, Morelia Michoacán, Mexico

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025

V. Bhateja et al. (eds.), *Innovations in ICT: Sustainability for Societal and Industrial Impact*, Lecture Notes in Networks and Systems 1364,
https://doi.org/10.1007/978-981-96-5220-4_23

group students into categories according to common patterns. A student who shows an affinity for mathematics and problem solving might be grouped into a set that suggests careers in fields such as engineering. This provides accurate, personalized guidance for students who are exploring their career options. Data science helps students understand the demands of the job market.

Keywords Educational guidance · Career choice · Automated learning · Clustering

1 Introduction

The economic growth of the regions and therefore of the labor market generates the need for better trained human resources to meet the needs of the public and private organizations that drive the economy of the countries. This need has contributed to the diversification of increasingly specialized professions and to the exponential emergence of new professional careers and occupations. In the same way, the requirements for education and training by formal education are greater and demand an occupational choice according to their skills. The socioeconomic and personal capabilities that guarantee job success and personal satisfaction, and better yet, that the choice is in line with the vocation built since childhood in the family. The above mentioned increases the difficulty for adolescent students to adequately and consciously choose their professional path; in addition to the psychosocial crisis, they are going through during the search and construction of their own identity and self-concept through existential questions that allow them to give meaning to their being and their existence. It is here where technology can be used, specifically automated learning, which through techniques such as clustering provides the conjunction of certain variables that would help the student to measure its scope and obtain a more favorable proposal for the development of his life in a professional and academic way. As more information is collected about themselves and their preferences, algorithms can adjust recommendations to provide more accurate and personalized guidance. As such, this research proposes using these technologies to work with counselors to provide personalized guidance, identify relevant career opportunities, and prepare students for a successful and rewarding future career. Currently, only talks and interviews with experts are offered, which allow students to learn about the context of the different areas in which they can help them select their careers, but they fall short and for many students they are not very helpful. It is therefore important to prepare the student's ground to introduce machine learning as a viable solution.

2 Materials and Methods

2.1 Orientation

It is a branch of education and psychology. According to Vidales in 1988, it distinguishes three types of orientation: (1) The professional, which addresses the world of work and productivity. (2) Vocational guidance is distinguished by the successful choice of the student's occupation and the extension of the term vocation, which includes the realization of the human being's capacity in his or her life. Finally, (3) educational or scholastic, which takes place in schools and aims at the educational success of individuals. This classification is based on the activities and areas that the vocational counselor performs and emphasizes. The "vocational guidance" is the one that is taught in the educational centers of higher middle level as a discipline, which focuses on studying the characteristics of the relationship of the human being with work, the continuation of studies, career and other activities of life such as hobbies, sports, and altruistic or philanthropic activities. It implies an integral process based on a humanistic vision that seeks the realization of man through a productive and, above all, satisfactory life. In this discipline, methods and techniques are used to study capabilities, values, motivations, and environmental factors important for making decisions related to work and occupation; and theories that explain the process of human choice at this stage of life. One of the two aspects observed by guidance is: (1) vocation, understood as a process of construction throughout life; and as a natural aspect of the individual, it is understood as a complex phenomenon constituted by beliefs, feelings, aspirations, and values that motivate the person to act in a certain direction; it is something subjective and internal to the individual, so that tools should be used as a technique to demonstrate the results and contribute to the guidance task closer to the contextual and social needs of the individual. And (2) that focuses on the counselor and refers to the observation of the development of the career and/or occupation; this is an objective and observable phenomenon that easily evidences its results. According to Vondracek et al. [33], they characterize vocational guidance and career development as the discipline that:

- It covers the entire life cycle and is related to the development of the person.
- It requires a multidisciplinary vision for its study.
- It must consider individual differences in career development, as well as the relationship of the individual to his or her environment.
- They study a wide variety of events related to decision making at different levels of analysis.

Education allows the development of countries, the growth of the economy, the increase of human capital and knowledge by its population, professions are a process that takes years of study and preparation in various areas, from health sciences, through exact sciences, business to arts, and sports. Virtually any area is an area of study, any aspiration to be a professional can be studied, and there is a path for those seeking to learn such knowledge and tools [30]. In the world, there are 7.67 billion

Table 1 Academic profile of the vocational counselor

Knowledge	• On the main theories of orientation • Factors influencing human development, motivational, and behavioral theories • Basic information on professions and job positions • Bases of assessment and measurement in psychology and education (psychometrics)
Skills	• Communication techniques • Psychometric evaluation methods • Interpretation of standardized and non-standardized tests • Management of didactic resources for work
Attitudes	• Acceptance of the person • Respect for the decisions, values, and aspirations of the individual • Positive consideration • Emotional maturity that allows him/her to work with the necessary knowledge • Responsibility and ethics

people, of which "only 6.7% of them have a university degree" according to the Independence University of the USA; even in more developed countries, the average number of higher education graduates is 45% of the population between 25–34 years old, according to the publication of the Organization for Economic Cooperation and Development [27]. In the USA alone, 46% of students have a college degree (Hanson 2021). In Mexico, on the other hand, only 24% of the population has this academic level [27].

2.2 Academic Profile of the Vocational Counselor

This discipline must be directed by a professional counselor, who tries to explain the factors that intervene in the development of a career and in the vocation of the individual. For this purpose, he/she must have a solid background in psychology, education, and human development, since he/she is a facilitator of decision making. This profile is around pedagogy, social work, psychology or, failing that, teacher educator if he/she complies with the following academic profile shown in Table 1.

2.3 Data Science

Data science is a multidisciplinary area that deals with processes and systems for extracting knowledge or insights from a large volume of data, which can be in different forms, structured or unstructured. This is a continuation of data analysis known as data mining or also known as knowledge discovery and data mining or KDD. Alex Liu (2015). The data science life cycle is divided into five stages: (1) collection, (2) preprocessing, (3) transformation, (4) data mining, (5) interpretation/

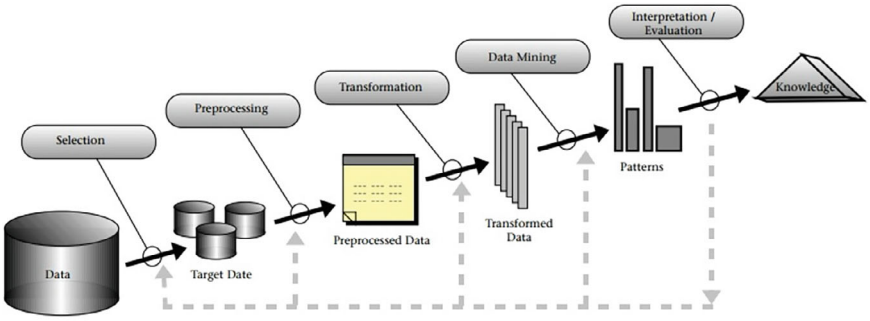

Fig. 1 Data science lifecycle. *Source* https://www.researchgate.net/publication/298352841_A_Model_For_Predicting_Ischemic_Stroke_Using_Data_Mining_Algorithms

evaluation, and (6) knowledge generation. Different tools or methodologies are used in each process, and it is the data scientist who is responsible for understanding them as shown in Fig. 1.

Kim [17] identifies three fundamental tools for data science which are (1) artificial intelligence, (2) machine learning, and (3) deep learning. He classifies these three tools as follows: deep learning is considered as a branch of machine learning, while machine learning is labeled as a branch of artificial intelligence. The same author defines the concept of machine learning as the process of building a model for data, where the data can be from any source (text, images, video, measurements, etc.), while the model constitutes the final product obtained by a machine learning algorithm. The dataset is usually divided into two subsets, called the training set and the test set. The training data is used to train the machine learning model, while the test data is used to validate it. This general scheme of dividing the dataset into two subsets, training the model and subjecting it to a validation process, is usually built iteratively until the result is obtained. This often helps to predict some event based on the analysis of the analyzed data. In some cases, it is the Contact Volume Editor that checks all the pdfs. In such cases, the authors are not involved in the checking phase.

2.4 Machine Learning (ML)

ML is one of the main branches of data science; "it is a form of artificial intelligence that allows a system to learn from data rather than learning through explicit programming" [10]. The normal ML process is that it ingests training data and within the model an output of information is produced once trained. These can be iterative and more accurate data-driven models can be produced [10]. There are different approaches and many techniques that can be applied to ML; these depend on the nature of the problem to be solved and the volume of data (big data). The main

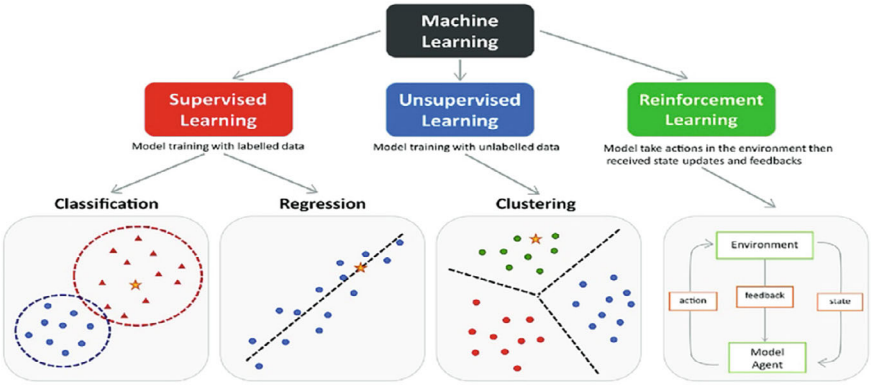

Fig. 2 Types of automated learning. *Source* https://www.interempresas.net

branches of ML are: (1) supervised learning, (2) unsupervised learning, (3) reinforcement learning, (4) semi-supervised learning, and (5) deep learning, as shown in Fig. 2.

Supervised learning has the characteristic of having an established dataset, and its focus is to find patterns. The data is characterized by a label that defines the meaning of the data. Unsupervised learning is characterized by unlabeled data. The algorithms classify the data based on the patterns or groupings found. Reinforcement learning is behavioral; this algorithm receives feedback from data analysis. The difference with supervised learning is that it does not take the training set and is more trial and error. There are several models used with machine learning, which are used to predict and solve problems. Clustering with K-means is an algorithm that is used to find groups of similar objects, objects that are more related to each other than objects in other groups (see Fig. 3). Examples of business-oriented clustering applications are the clustering of documents, music, and movies by distinct topics based on common shopping behaviors as a basis for recommendation engines. Clustering groups similar data. Although it is not a prediction model, it allows to run data classification, Ikotun et al. (2023).

According to Martinez et al. (2023), artificial neural networks (ANNs) are 9) inspired by the structure of the human brain, deep neural networks are used for complex tasks such as image processing and natural language processing. ANNs replace the functioning of the brain in a computer. These are just a few examples, and there are many other models and algorithms used in machine learning. The choice of model depends on the specific problem and the available data. Finally, deep learning is a very specific type of ML because it involves a solution based on artificial neural networks, which push a gradient through a network with millions of weights, i.e., through different iterative layers. The artificial intelligence classification is very useful for data type solutions such as imaging, speech, and computer vision applications.

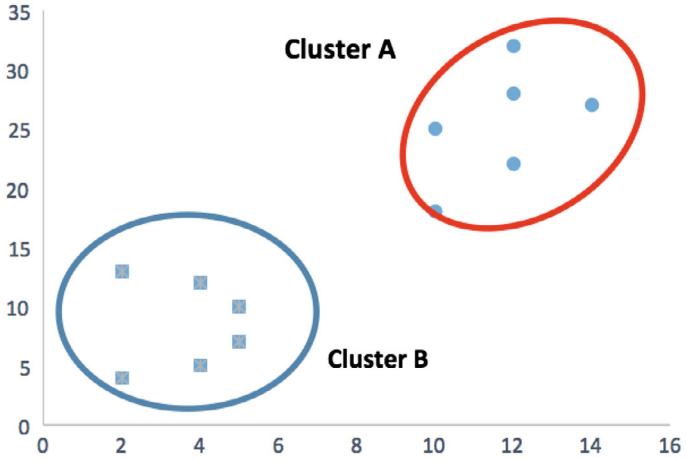

Fig. 3 K-means, clustering. *Source* https://octo.vmware.com/3-approaches-to-machine-learning/

2.5 Machine Learning Interpretability

It is complicated to explain interpretability mathematically, so, on this occasion, from a "non-mathematical" perspective, interpretability is defined as the degree to which a human can understand the cause of a decision (Miller, 2019). For their part, Kim et al. [16] refer that interpretability is the degree to which a human can consistently predict the outcome of the model. It can be said then, that, the higher the interpretability of an ML model, the easier it is for someone to understand why certain decisions have been made or certain predictions have been made [25]. One model is more interpretable than another if its decisions are easier for a human to understand than the decisions of the other model. Consideration should be given to:

- Scope of interpretability. An algorithm trains a model that makes predictions; each step of the model can be evaluated in terms of transparency interpretability.
- The transparency of the algorithm refers to how a model learns from the (input) data and what kind of relationships it can recognize.
- Algorithms such as the least squares method for linear regression models are well studied and understood. They are characterized by high transparency.
- Deep learning approaches are not as well understood, and their inner workings are the subject of ongoing research. The latter are considered less transparent [25].
- Interpretable model. A model could be described as interpretable if the whole model can be understood at once [20].
- To explain the overall outcome of the model, the trained model, knowledge of the algorithm, and the data are needed. According to Molnar [25], this level of interpretability is about understanding how the model makes decisions, based on which features are important and what kind of interactions occur between them.

2.6 Public Policy Guidelines for Secondary Education

The Mexican Ministry of Public Education presented in 2019 the "Lines of Public Policies for Secondary Education"; where six general topics are addressed: (1) quality and equity; (2) contents and activities for learning; (3) dignification of teachers and revaluation of working conditions; (4) school governance; (5) educational infrastructure; and (6) financing and resources. The axis of education with quality and equity establishes that an urgent problem to be solved is school dropout. It is proposed to review the conditions of the student community at this educational level, as well as the functioning of the programs implemented to reduce school dropout. The section on educational infrastructure focuses on improving the physical conditions and services offered in educational facilities to ensure an environment conducive to learning; and, among others, it proposes the following: "Strengthening vocational and psychological support services for students." Educational guidance contributes to what is established in the two previous sections, and to fulfill its purpose, it must be planned, programmed, executed, and evaluated [12]. Once the problem has been identified with first year high school students, the levels of knowledge they have about the context of their choice must be measured. These levels are stored in a dataset, which allows their subsequent analysis through data science. It was defined that the sustenance and scope of the research is grounded through different approaches among which are the approach: (1) experimental, (2) exploratory, (3) descriptive and (4) correlational [18]. For the development of the research, a set of stages were defined, which are shown in Fig. 4.

2.7 Educational Guidance Program to Help High School Students Decide on a Professional Career

The educational guidance program seeks to favor vocational decision making in high school students, intervening in a timely manner in the initial stage of this educational level, considering that most of them were not provided with any kind of vocational support by the school institutions to which they have belonged. This work will support the vocational choice in a conscious way, trying to stick to the capabilities, interests, economic, and social needs of the context in which it is intended to develop. In general, three levels of intervention are proposed, seeking to stimulate certain thinking actions. (1) Level I corresponds to the learning categories observed in knowledge and comprehension thinking actions of Bloom's taxonomy proposed in 1956, which serves as the theoretical basis for the competency model, (2) level II makes use of the level of analysis and synthesis of information, and 3) level III refers to the level of application and evaluation of the same taxonomy, aiming to lead the student to the knowledge of information so that he/she is aware of the current situation and of his/her role as a social individual who contributes with his/her argued and solid decisions to the world of knowledge and innovation in the globalized world

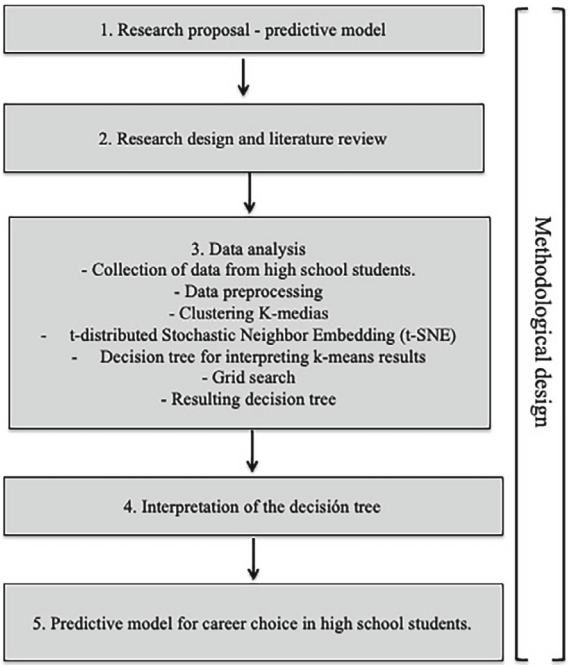

Fig. 4 Stages of the investigation. *Source* own

in which he/she finds himself/herself. In measuring instrument level I, measurement can be defined according to Carmines and Zeller (1979) as the process of linking abstract processes with empirical indicators through an explicit and organized plan to classify. In this process, the measurement or data collection instrument plays a central role since; without it, there are no classified observations. The measurement instrument proposed to be applied before the development of the intervention program (pre-test) is the survey called "Analysis of the need for Vocational Guidance in High School students," by Guillén [9] and which employs a Likert [19] proposal, designed in the electronic platform e-encuesta (www.e-encuesta.com), made up of 24 questions related to the variables of the research, and which can be accessed at the following electronic address: http://www.eencuesta.com/s/MVKJlD7SQnuwt9wm2xSzwg/. The instrument designed for the pre-test phase contributes to obtaining the following data: (1) Conduct a diagnosis on the knowledge of the subject by the population to be studied. (2) To have a broad knowledge of the population under study. (3) Detect the real need on vocational aspects. (4) Locate respondents in socioeconomic level; know their school environment, parental influence.

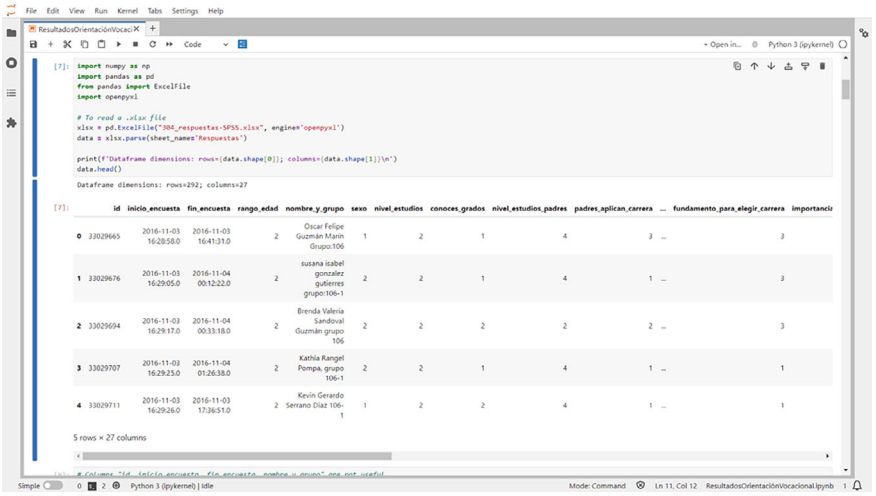

Fig. 5 Dataset import process. *Source* own

3 Results

3.1 Data Analysis Using ML

The dataset obtained from the survey applied to the students enrolled in the educational guidance program does not contain response variables to make predictions, nor labels to classify. The instructions for importing the dataset are shown in Fig. 5.

With these restrictions, the clustering technique is appropriate for exploring the relationships between features. Unlike supervised classification, in the absence of class labels, the objective is to group objects into clusters according to their observable characteristics; so that each cluster contains similar objects, and others different. As mentioned by MacQueen [21]. The K-means method provides a single partition of the data into a fixed number of clusters; and nowadays, it has been improved with new initialization methods [1]. In another sense, one of the main directions of ML is to understand the reasoning behind the decisions made by the models [20, 25, 26]. Traditional clustering methods, including K-means, assign points to clusters based on some distance metric, without providing any specific description of the clusters and the reasons why points are assigned to them. However, for this proposal, it is critical to understand how the resulting clusters differ, whether by demographics, diagnostics, or other factors. Although the importance of cluster interpretability is beginning to gain traction, there has been limited success in addressing this issue [8]. A common approach is the visualization of clusters in a two-dimensional graph using principal component analysis (PCA) projections [13, 28]; by reducing the dimensionality of the feature space, PCA renders the relationship between the clusters and the original variables opaque. On the other hand, for the dimensional reduction

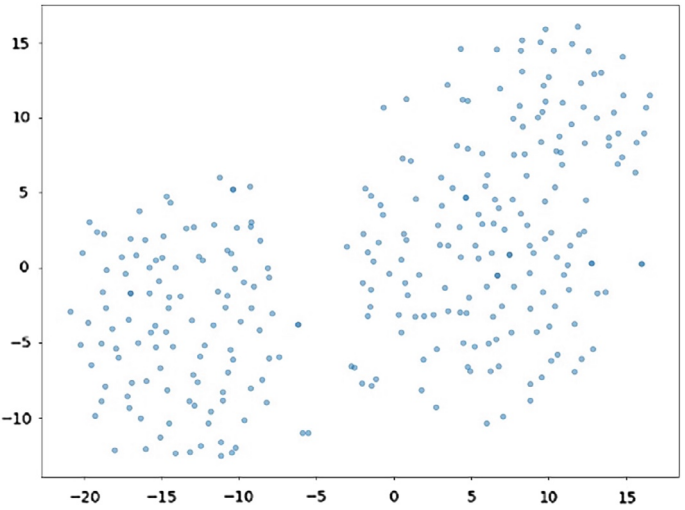

Fig. 6 t-SNE visualization of the pre-test dataset. *Source* own

stage, it is proposed to use the t-distributed Stochastic Neighbor Embedding (t-SNE) technique, which is a technique for the visualization of high dimensional data, developed by Hinton and Van Der Maaten [31]. As in the case of the PCA, the objective is to determine one more space of smaller dimension always keeping the same distance between points, improving interpretability. Figure 6 shows the clusters obtained from the application of t-SNE.

Tree-based supervised learning methods, such as Classification and Regression Trees (CART) [5], are naturally suited to problems that prioritize interpretability, as their feature splits and decision paths provide insight into the differences between variables. Taking the above into consideration, the Interpretable Clustering via Optimal Trees (ICOT) algorithm, developed by Bertsimas and Dunn [3] is chosen to interpret the results of the educational guidance program intervention level I measurement using a K-means model, subsequently fitting an Optimal Classification Tree (OCT) to predict the number of cluster assigned to each sample in our dataset by the K-means model. In this way, the resulting tree will provide, without manual intervention, interpretations of the clusters and insights into how points are assigned to these clusters.

4 Discussion and Conclusions

This model for measuring, evaluating, and describing the need and relevance of the educational guidance program for high school students through interpretable clustering is fundamental for understanding the effectiveness and relevance of such

a program in the educational context. The use of interpretive clustering techniques allows identifying meaningful patterns in the data collected, which facilitates a deeper understanding of students' needs and characteristics. In addition, this approach can help to customize educational guidance interventions, adapting them to the different needs and profiles of students. The evaluation of program effectiveness through this model provides valuable information to improve its design and implementation, as well as to identify areas for improvement in the educational guidance offered to high school students. The application of interpretable clustering techniques in this context provides a powerful tool to understand and improve the educational guidance program, thus contributing to the integral development of students during this crucial stage of their academic formation. Educational guidance in high school helps students develop strong self-knowledge and identify careers or trades that best match their strengths and preferences, which increases the likelihood of long-term academic and career success. In addition, one-on-one support from a guidance counselor promotes a sense of security and emotional well-being in teens, which can help reduce dropout rates. Machine learning, and specifically, the clustering model, can help identify the relevance and need for the educational guidance program, through interpretability. It is important to identify the model to be applied to the dataset; it will determine whether classification, clustering, linear regression of data or dimension reduction is required. The proposed model for measuring, evaluating, and describing the need and relevance of the educational guidance program for first year high school students through interpretable clustering allows focusing resources on the student segment susceptible to dropping out. In addition to identifying through the clustering model that the two main groups are: (1) those students who have the fundamentals and elements necessary to choose a career and (2) those who have the objective of dedicating themselves to a trade, since this has been traditionally repeated for generations in their families.

References

1. Arthur D, Vassilvitskii S (2007) k-means++: The advantages of careful seeding. In: Proceedings of the 18th annual ACM-SIAM symposium on discrete algorithms. Society for Industrial and Applied Mathematics, pp 1027–1035
2. Berkley Ude (2018) What is Data Science? https://ischoolonline.berkeley.edu/data-science/what-is-data-science/
3. Bertsimas D, Dunn J (2019) Machine learning under a modern optimization lens. Dynamic Ideas Press, Waltham
4. Bertsimas D, Orfanoudaki A, Wiberg H (2020) Interpretable clustering: and optimization approach. Mach Learn 110(1):89–138. https://doi.org/10.1007/s10994-020-05896-2
5. Breiedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. CRC Press, Boca Raton
6. Beckers S, Chockler H, Halpern JY (2022) A causal analysis of harm. In: Beckers S, Chockler H (eds) Proceedings of the 36th conference on neural information processing systems (NeurIPS), November
7. Carmines EG, Zeller RA (1979) Reliability and validity assessment. vol 17.

8. Doshi VF, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1702.08608
9. Guillén Y (2018) Programa para favorecer la toma de decisión de una carrera profesional en estudiantes de Bachillerato [Master's thesis, Universidad Autónoma de México]. https://hdl.handle.net/20.500.14330/TES01000775717
10. IBM (2021) Machine Learning. Retrieved from https://www.ibm.com/mx-es/analytics/machine-learning
11. Ikotun AM, Ezugwu AE, Abualigah L, Abuhaija B, Heming J (2023) K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data. Inf Sci 622:178–210
12. Jarrin P (2002) In: Vocational and professional development. Don Bosco, Quito
13. Jolliffe I (2011) Principal component analysis. In: International encyclopedia of statistical science. Springer, pp 1094–1096
14. Juárez C (2022) Multiple linear regression. Retrieved from Multiple regression and other multivariate techniques: http://sigma.iimas.unam.mx/claudiajrz/ej1multiple.html
15. Khalid B, Abdelwahab N (2015) A model for predicting Ischemic stroke using data mining algorithms. Int J Innov Sci Eng Technol 2(11):18–23
16. Kim B, Khanna R, Koyejo O (2016) Examples are not enough, learn to criticize! criticism for interpretability. In: Proceedings of the 30th international conference on neural information processing systems. pp 2288–2296. Retrieved from. https://papers.nips.cc/paper_files/paper/2016/hash/5680522b8e2bb01943234bce7bf84534-Abstract.html
17. Kim P (2017) In: MATLAB deep learning. Apress. https://doi.org/10.1007/978-1-4842-2845-6
18. Kothari C, Garg G (2014) Research methodology methods and techniques. 3rd ed. New Delhi, New Age International
19. Likert R (1932) A technique for the measurement of attitudes. Archives of Psychol 22(140):55
20. Lipton ZC (2016) The mythos of model interpretability. arXiv (CornellUniversity). https://doi.org/10.48550/arxiv.1606.03490
21. MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley symposium on mathematical statistics and probability, volume 1: Statistics. Berkeley, CA, University of California Press, pp 281–297
22. Martínez CM, Rigueira DX, Larrañaga JA, Martínez TJ, Ocarranza PI, Kreibel D (2023) Impact of artificial intelligence on assessment methods in primary and secondary education: systematic review of the literature. J Psychodidactics
23. Metalworking (2022) Defect monitoring and machine learning in L-PBF additive technology. Retrieved from Metalmechanics: https://www.interempresas.net/MetalMecanica/Articulos/385970-Monitorizacion-de-defectos-y-machine-learning-en-la-tecnologia-aditiva-L-PBF.html
24. Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38. https://doi.org/10.1016/j.artint.2018.07.007
25. Molnar C (2022) Interpretable machine learning: a guide for making black box models explainable (2nd ed.). https://christophm.github.io/interpretable-ml-book/
26. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B (2019) Definitions, methods, and applications in interpretable machine learning. Proc National Acad Sci 116(44):22071–22080. https://doi.org/10.1073/pnas.1900654116
27. OECD (2020) Hours worked. https://data.oecd.org/emp/hours-worked.htm
28. Rao CR (1964) The use and interpretation of principal component analysis in applied research. Sankhyā: The Indian J Stat Series A (1961–2002) 26(4):329–358
29. Raschka S, Mirjalilli V (2019) Python machine learning: machine learning and deep learning with python, Scikit-Learn and TensorFlow (2.a ed.). Sage Publications
30. Sánchez Escobedo PA, Valdés Cuervo ÁA (2003) Teoría y práctica de la orientación en la escuela: Un enfoque psicológica. D.F., Mexico, Manual Moderno
31. Van Der Maaten L, Hinton GE (2008) Visualizing Data using t-SNE. J Mach Learn Res 9(86):2579–2605. http://isplab.tudelft.nl/sites/default/files/vandermaaten08a.pdf

32. Vidales I (1988) Nuevas prácticas de Orientación Vocacional. Editorial Trillas. Mexico, pp 99–103
33. Vondracek FW, Schulenberg JE (1986) Career development in adolescence: some conceptual and intervention issues. The Vocational Guidance Quarter 34:247–254