Andrew M. Olney · Irene-Angelica Chounta · Zitao Liu · Olga C. Santos · Ig Ibert Bittencourt (Eds.)

LNAI 14829

Artificial Intelligence in Education

25th International Conference, AIED 2024
Recife, Brazil, July 8–12, 2024
Proceedings, Part I

Part I

AIED RECIFE 2024
25th International Conference on Artificial Intelligence in Education

Springer

Lecture Notes in Computer Science

Lecture Notes in Artificial Intelligence 14829

Founding Editor

Jörg Siekmann

The series Lecture Notes in Artificial Intelligence (LNAI) was established in 1988 as a topical subseries of LNCS devoted to artificial intelligence.

The series publishes state-of-the-art research results at a high level. As with the LNCS mother series, the mission of the series is to serve the international R & D community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings.

Andrew M. Olney · Irene-Angelica Chounta ·
Zitao Liu · Olga C. Santos · Ig Ibert Bittencourt
Editors

Artificial Intelligence in Education

25th International Conference, AIED 2024
Recife, Brazil, July 8–12, 2024
Proceedings, Part I

Springer

Editors
Andrew M. Olney
University of Memphis
Memphis, TN, USA

Irene-Angelica Chounta
University of Duisburg-Essen
Duisburg, Germany

Zitao Liu
Jinan University
Guangzhou, China

Olga C. Santos
UNED
Madrid, Spain

Ig Ibert Bittencourt
Universidade Federal de Alagoas
Maceio, Brazil

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-031-64301-9 ISBN 978-3-031-64302-6 (eBook)
https://doi.org/10.1007/978-3-031-64302-6

LNCS Sublibrary: SL7 – Artificial Intelligence

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The 25th International Conference on Artificial Intelligence in Education (AIED 2024) was hosted by Centro de Estudos e Sistemas Avançados do Recife (CESAR), Brazil from July 8 to July 12, 2024. It was set up in a face-to-face format but included an option for an online audience. AIED 2024 was the next in a longstanding series of annual international conferences for the presentation of high-quality research on intelligent systems and the cognitive sciences for the improvement and advancement of education. Note that AIED is ranked A in CORE (top 16% of all 783 ranked venues), the well-known ranking of computer science conferences. The AIED conferences are organized by the prestigious International Artificial Intelligence in Education Society, a global association of researchers and academics, which has already celebrated its 30th anniversary, and aims to advance the science and engineering of intelligent human-technology ecosystems that support learning by promoting rigorous research and development of interactive and adaptive learning environments for learners of all ages across all domains.

The theme for the AIED 2024 conference was "AIED for a World in Transition". The conference aimed to explore how AI can be used to enhance the learning experiences of students and teachers alike when disruptive technologies are turning education upside down. Rapid advances in Artificial Intelligence (AI) have created opportunities not only for personalized and immersive experiences but also for ad hoc learning by engaging with cutting-edge technology continually, extending classroom borders, from engaging in real-time conversations with large language models (LLMs) to creating expressive artifacts such as digital images with generative AI or physically interacting with the environment for a more embodied learning. As a result, we now need new approaches and measurements to harness this potential and ensure that we can safely and responsibly cope with a world in transition. The conference seeks to stimulate discussion of how AI can shape education for all sectors, how to advance the science and engineering of AI-assisted learning systems, and how to promote broad adoption.

AIED 2024 attracted broad participation. We received 334 submissions for the main program, of which 280 were submitted as full papers, and 54 were submitted as short papers. Of the full paper submissions, 49 were accepted as full papers, and another 27 were accepted as short papers. The acceptance rate for full papers and short papers together was 23%. These accepted contributions are published in the Springer proceedings volumes LNAI 14829 and 14830.

The submissions underwent a rigorous double-masked peer-review process aimed to reduce evaluation bias as much as possible. The first step of the review process was done by the program chairs, who verified that all papers were appropriate for the conference and properly anonymized. Program committee members were asked to declare conflicts of interest. After the initial revision, the program committee members were invited to bid on the anonymized papers that were not in conflict according to their declared conflicts of interest. With this information, the program chairs made the review assignment, which consisted of three regular members to review each paper plus a senior member to provide a meta-review. The management of the review process (i.e., bidding, assignment, discussion, and meta-review) was done with the EasyChair platform, which was configured so that reviewers of the same paper were anonymous to each other. A subset of the program committee members were not included in the initial assignment but were asked to be ready to do reviews that were not submitted on time (i.e., the emergency review period). To avoid a situation where program committee members would be involved in too many submissions, we balanced review assignments and then rebalanced them during the emergency review period.

As a result, each submission was reviewed anonymously by at least three Program Committee (PC) members and then a discussion was led by a Senior Program Committee (SPC) member. PC and SPC members were selected based on their authorship in previous AIED conferences, their experience as reviewers in previous AIED editions, their h-index as calculated by Google Scholar, and their previous positions in organizing and reviewing related conferences. Therefore all members were active researchers in the field, and SPC members were particularly accomplished on these metrics. SPC members served as meta-reviewers whose role was to seek consensus to reach the final decision about acceptance and to provide the corresponding meta-review. They were also asked to check and highlight any possible biases or inappropriate reviews. Decisions to accept/reject were taken by the program chairs. For borderline cases, the contents of the paper were read in detail before reaching the final decision. In summary, we are confident that the review process assured a fair and equal evaluation of the submissions received without any bias, as far as we are aware.

Beyond paper presentations, the conference included a Doctoral Consortium Track, Late-Breaking Results, a Workshops and Tutorials Track, and an Industry, Innovation and Practitioner Track. There was a WideAIED track, which was established in 2023, where opportunities and challenges of AI in education were discussed with a global perspective and with contributions coming also from areas of the world that are currently under-represented in AIED. Additionally, a BlueSky special track was included with contributions that reflect upon the progress of AIED so far and envision what is to come in the future. The submissions for all these tracks underwent a rigorous peer review process. Each submission was reviewed by at least three members of the AIED community, assigned by the corresponding track organizers who then took the final decision about acceptance.

The participants of the conference had the opportunity to attend three keynote talks: “Navigating Strategic Challenges in Education in the Post-Pandemic AI Era” by Blaženka Divjak, “Navigating the Evolution: The Rising Tide of Large Language Models for AI and Education” by Peter Clark, and “Artificial Intelligence in Education and Public Policy: A Case from Brazil” by Seiji Isotani. These contributions are published in the Springer proceedings volumes CCIS 2150 and 2151.

The conference also included a Panel with experts in the field and the opportunity for the participants to present a demonstration of their AIED system in a specific session of Interactive Events. A selection of the systems presented is included as showcases on the web page of the IAIED Society[1]. Finally, there was a session with presentations

[1] https://iaied.org/showcase.

of papers published in the International Journal of Artificial Intelligence in Education[2], the journal of the IAIED Society indexed in the main databases, and a session with the best papers from conferences of the International Alliance to Advance Learning in the Digital Era (IAALDE)[3], an alliance of research societies that focus on advances in computer-supported learning, to which the IAIED Society belongs.

For making AIED 2024 possible, we thank the AIED 2024 Organizing Committee, the hundreds of Program Committee members, the Senior Program Committee members, and the AIED proceedings chairs Paraskevi Topali and Rafael D. Araújo. In addition, we would like to thank the Executive Committee of the IAIED Society for their advice during the conference preparation, and specifically two of the working groups, the Conference Steering Committee, and the Diversity and Inclusion working group. They all gave their time and expertise generously and helped with shaping a stimulating AIED 2024 conference. We are extremely grateful to everyone!

July 2024

Andrew M. Olney
Irene-Angelica Chounta
Zitao Liu
Olga C. Santos
Ig Ibert Bittencourt

[2] https://link.springer.com/journal/40593.
[3] https://alliancelss.com/.

Organization

Conference General Co-chairs

Olga C. Santos	UNED, Spain
Ig Ibert Bittencourt	Universidade Federal de Alagoas, Brazil

Program Co-chairs

Andrew M. Olney	University of Memphis, USA
Irene-Angelica Chounta	University of Duisburg-Essen, Germany
Zitao Liu	Jinan University, China

Doctoral Consortium Co-chairs

Yu Lu	Beijing Normal University, China
Elaine Harada T. Oliveira	Universidade Federal do Amazonas, Brazil
Vanda Luengo	Sorbonne Université, France

Workshop and Tutorials Co-chairs

Cristian Cechinel	Federal University of Santa Catarina, Brazil
Carrie Demmans Epp	University of Alberta, Canada

Interactive Events Co-chairs

Leonardo B. Marques	Federal University of Alagoas, Brazil
Ben Nye	University of Southern California, USA
Rwitajit Majumdar	Kumamoto University, Japan

Industry, Innovation and Practitioner Co-chairs

Diego Dermeval	Federal University of Alagoas, Brazil
Richard Tong	IEEE Artificial Intelligence Standards Committee, USA
Sreecharan Sankaranarayanan	Amazon, USA
Insa Reichow	German Research Center for Artificial Intelligence, Germany

Posters and Late Breaking Results Co-chairs

Marie-Luce Bourguet	Queen Mary University of London, UK
Qianru Liang	Jinan University, China
Jingyun Wang	Durham University, UK

Panel Co-chairs

Julita Vassileva	University of Saskatchewan, Canada
Alexandra Cristea	Durham University, UK

Blue Sky Co-chairs

Ryan S. Baker	University of Pennsylvania, USA
Benedict du Boulay	University of Sussex, UK
Mirko Marras	University of Cagliari, Italy

WideAIED Co-chairs

Isabel Hilliger	Pontificia Universidad Católica, Chile
Marco Temperini	Sapienza University of Rome, Italy
Ifeoma Adaji	University of British Columbia, Canada
Maomi Ueno	University of Electro-Communications, Japan

Local Organising Co-chairs

Rafael Ferreira Mello	Universidade Federal Rural de Pernambuco, Brazil
Taciana Pontual	Universidade Federal Rural de Pernambuco, Brazil

AIED Mentoring Fellowship Co-chairs

Amruth N. Kumar	Ramapo College of New Jersey, USA
Vania Dimitrova	University of Leeds, UK

Diversity and Inclusion Co-chairs

Rod Roscoe	Arizona State University, USA
Kaska Porayska-Pomsta	University College London, UK

Virtual Experiences Co-chairs

Guanliang Chen	Monash University, Australia
Teng Guo	Jinan University, China
Eduardo A. Oliveira	University of Melbourne, Australia

Publicity Co-chairs

Son T. H. Pham	Nha Viet Institute, USA
Pham Duc Tho	Hung Vuong University, Vietnam
Miguel Portaz	UNED, Spain

Volunteer Co-chairs

Isabela Gasparini	Santa Catarina State University, Brazil
Lele Sha	Monash University, Australia

Proceedings Co-chairs

Paraskevi Topali	Radboud University, Netherlands
Rafael D. Araújo	Federal University of Uberlândia, Brazil

Awards Co-chairs

Ning Wang	University of Southern California, USA
Beverly Woolf	University of Massachusetts, USA

Sponsorship Chair

Tanci Simões Gomes	CESAR, Brazil

Scholarship Chair

Patrícia Tedesco	Universidade Federal de Pernambuco, Brazil

Steering Committee

Noboru Matsuda	North Carolina State University, USA
Eva Millan	Universidad de Málaga, Spain
Sergey Sosnovsky	Utrecht University, Netherlands
Ido Roll	Israel Institute of Technology, Israel
Maria Mercedes T. Rodrigo	Ateneo de Manila University, Philippines

Senior Program Committee Members

Laura Allen	University of Minnesota, USA
Ryan Baker	University of Pennsylvania, USA
Gautam Biswas	Vanderbilt University, USA
Nigel Bosch	University of Illinois at Urbana-Champaign, USA
Jesus G. Boticario	UNED, Spain
Bert Bredeweg	University of Amsterdam, Netherlands
Christopher Brooks	University of Michigan, USA
Guanliang Chen	Monash University, Australia

Ruth Cobos	Universidad Autónoma de Madrid, Spain
Cesar A. Collazos	Universidad del Cauca, Colombia
Mutlu Cukurova	University College London, UK
Maria Cutumisu	McGill University, Canada
Tenzin Doleck	Simon Fraser University, Canada
Ralph Ewerth	Leibniz Universität Hannover, Germany
Dragan Gasevic	Monash University, Australia
Isabela Gasparini	UDESC, Brazil
Sébastien George	Le Mans Université, France
Alex Sandro Gomes	Universidade Federal de Pernambuco, Brazil
Jason Harley	McGill University, Canada
Peter Hastings	DePaul University, USA
Bastiaan Heeren	Open University, Netherlands
Neil Heffernan	Worcester Polytechnic Institute, USA
Tomoo Inoue	University of Tsukuba, Japan
Seiji Isotani	Harvard University, USA
Johan Jeuring	Utrecht University, Netherlands
Srecko Joksimovic	University of South Australia, Australia
Simon Knight	UTS, Australia
Kenneth Koedinger	Carnegie Mellon University, USA
Irena Koprinska	University of Sydney, Australia
Vitomir Kovanovic	University of South Australia, Australia
Amruth N. Kumar	Ramapo College of New Jersey, USA
H. Chad Lane	University of Illinois at Urbana-Champaign, USA
James Lester	North Carolina State University, USA
Fuhua Lin	Athabasca University, Canada
Roberto Martinez-Maldonado	Monash University, Australia
Riichiro Mizoguchi	Japan Advanced Institute of Science and Technology, Japan
Kasia Muldner	Carleton University, Canada
Tomohiro Nagashima	Saarland University, Germany
Roger Nkambou	Université du Québec à Montréal, Canada
Radek Pelánek	Masaryk University Brno, Czechia
Elvira Popescu	University of Craiova, Romania
Kaska Porayska-Pomsta	University College London, UK
Maria Mercedes T. Rodrigo	Ateneo de Manila University, Philippines
Jonathan Rowe	North Carolina State University, USA
Demetrios Sampson	Curtin University, Australia
Mohammed Saqr	University of Eastern Finland, Finland
Filippo Sciarrone	Universitas Mercatorum, Italy
Sergey Sosnovsky	Utrecht University, Netherlands
Michelle Taub	University of Central Florida, USA

Craig Thompson	University of British Columbia, Canada
Stefan Trausan-Matu	University Politehnica of Bucharest, Romania
Maomi Ueno	University of Electro-Communications, Japan
Rosa Vicari	Universidade Federal do Rio Grande do Sul, Brazil
Vincent Wade	Trinity College Dublin, Ireland
Alistair Willis	Open University, UK
Diego Zapata-Rivera	Educational Testing Service, USA
Gustavo Zurita	Universidad de Chile, Chile

Program Committee Members

Mark Abdelshiheed	University of Colorado Boulder, USA
Kamil Akhuseyinoglu	University of Pittsburgh, USA
Carlos Alario-Hoyos	Universidad Carlos III de Madrid, Spain
Laia Albó	Universitat Pompeu Fabra, Spain
Giora Alexandron	Weizmann Institute of Science, Israel
Isaac Alpizar Chacon	Utrecht University, Netherlands
Ioannis Anastasopoulos	University of California, Berkeley, USA
Antonio R. Anaya	Universidad Nacional de Educación a Distancia, Spain
Tracy Arner	Arizona State University, USA
Burcu Arslan	Educational Testing Service, USA
Juan I. Asensio-Pérez	Universidad de Valladolid, Spain
Michelle Banawan	Asian Institute of Management, Philippines
Abhinava Barthakur	University of South Australia, Australia
Beata Beigman Klebanov	Educational Testing Service, USA
Brian Belland	Pennsylvania State University, USA
Francisco Bellas	Universidade da Coruna, Spain
Emmanuel Blanchard	Le Mans Université, France
Nathaniel Blanchard	Colorado State University, USA
Maria Bolsinova	Tilburg University, Netherlands
Miguel L. Bote-Lorenzo	Universidad de Valladolid, Spain
Anthony F. Botelho	University of Florida, USA
François Bouchet	Sorbonne Université - LIP6, France
Marie-Luce Bourguet	Queen Mary University of London, UK
Rex Bringula	University of the East, Philippines
Julien Broisin	Université Toulouse III - Paul Sabatier, France
Armelle Brun	LORIA - Université de Lorraine, France
Minghao Cai	University of Alberta, Canada
Dan Carpenter	North Carolina State University, USA

R. McKell Carter	University of Colorado Boulder, USA
Paulo Carvalho	Carnegie Mellon University, USA
Alberto Casas	UNED, Spain
Teresa Cerratto-Pargman	Stockholm University, Sweden
Cs Chai	Chinese University of Hong Kong, China
Jiahao Chen	TAL, China
Penghe Chen	Beijing Normal University, China
Thomas K. F. Chiu	Chinese University of Hong Kong, China
Heeryung Choi	Massachusetts Institute of Technology, USA
Chih-Yueh Chou	元智大工程系, Taiwan
Jody Clarke-Midura	Utah State University, USA
Keith Cochran	DePaul University, USA
Maria de Los Angeles Constantino González	Tecnológico de Monterrey Campus Laguna, Mexico
Evandro Costa	Federal University of Alagoas, Brazil
Alexandra Cristea	Durham University, UK
Jeffrey Cross	Tokyo Institute of Technology, Japan
Mihai Dascalu	University Politehnica of Bucharest, Romania
Jeanine DeFalco	University of New Haven, USA
Carrie Demmans Epp	University of Alberta, Canada
Vanessa Dennen	Florida State University, USA
Michel Desmarais	Polytechnique Montréal, Canada
M. Ali Akber Dewan	Athabasca University, Canada
Yannis Dimitriadis	University of Valladolid, Spain
Konomu Dobashi	Aichi University, Japan
Fabiano Dorça	Universidade Federal de Uberlandia, Brazil
Mohsen Dorodchi	University of North Carolina Charlotte, USA
Benedict du Boulay	University of Sussex, UK
Cristina Dumdumaya	University of Southeastern Philippines, Philippines
Nicholas Duran	Arizona State University, USA
Kareem Edouard	Drexel University, USA
Fahmid Morshed Fahid	North Carolina State University, USA
Xiuyi Fan	Nanyang Technological University, Singapore
Alexandra Farazouli	Stockholm University, Sweden
Effat Farhana	Vanderbilt University, USA
Mingyu Feng	WestEd, USA
Márcia Fernandes	Federal University of Uberlândia, Brazil
Brendan Flanagan	Kyoto University, Japan
Carol Forsyth	Educational Testing Service, USA
Reva Freedman	Northern Illinois University, USA

Cristiano Galafassi	Universidade Federal do Rio Grande do Sul, Brazil
Wenbin Gan	National Institute of Information and Communications Technology, Japan
Zhen Gao	McMaster University, Canada
Yiannis Georgiou	Cyprus University of Technology, Cyprus
Alireza Gharahighehi	KU Leuven, Belgium
Lucia Giraffa	Pontificia Universidade Católica do Rio Grande do Sul, Brazil
Michael Glass	Valparaiso University, USA
Benjamin Goldberg	United States Army DEVCOM Soldier Center, USA
Aldo Gordillo	Universidad Politécnica de Madrid, Spain
Guher Gorgun	University of Alberta, Canada
Cyril Goutte	National Research Council Canada, Canada
Sabine Graf	Athabasca University, Canada
Monique Grandbastien	LORIA, Université de Lorraine, France
Beate Grawemeyer	Coventry University, UK
Quanlong Guan	Jinan University, China
Nathalie Guin	LIRIS - Université de Lyon, France
Ella Haig	University of Portsmouth, UK
Jiangang Hao	Educational Testing Service, USA
Yugo Hayashi	Ritsumeikan University, Japan
Yusuke Hayashi	Hiroshima University, Japan
Arto Hellas	Aalto University, Finland
Davinia Hernandez-Leo	Universitat Pompeu Fabra, Barcelona, Spain
Anett Hoppe	Leibniz Universität Hannover, Germany
Iris Howley	Williams College, USA
Sharon Hsiao	Santa Clara University, USA
Xiangen Hu	University of Memphis, USA
Lingyun Huang	Education University of Hong Kong, China
Shuyan Huang	TAL Education Group, China
Stephen Hutt	University of Denver, USA
Tsunenori Ishioka	National Center for University Entrance Examinations, Japan
Patricia Jaques	UFPEL, UFPR, Brazil
Stéphanie Jean-Daubias	Université de Lyon, LIRIS, France
Yang Jiang	Columbia University, USA
Yueqiao Jin	Monash University, Australia
David Joyner	Georgia Institute of Technology, USA
Hamid Karimi	Utah State University, USA
Tanja Käser	EPFL, Switzerland

Akihiro Kashihara	University of Electro-Communications, Japan
Enkelejda Kasneci	Technical University of Munich, Germany
Judy Kay	University of Sydney, Australia
Mizue Kayama	Shinshu University, Japan
Mohammad Khalil	University of Bergen, Norway
Hassan Khosravi	University of Queensland, Australia
Chanmin Kim	Pennsylvania State University, USA
Yeo Jin Kim	North Carolina State University, USA
Yoon Jeon Kim	University of Wisconsin-Madison, USA
Rene Kizilcec	Cornell University, USA
Elizabeth Koh	Nanyang Technological University, Singapore
Kazuaki Kojima	Teikyo University, Japan
Sotiris Kotsiantis	University of Patras, Greece
Eleni Kyza	Cyprus University of Technology, Cyprus
Sébastien Lallé	Sorbonne University, France
Andrew Lan	University of Massachusetts Amherst, USA
Mikel Larrañaga	University of the Basque Country, Spain
Elise Lavoué	Université Jean Moulin Lyon 3, LIRIS, France
Nguyen-Thinh Le	Humboldt Universität zu Berlin, Germany
Tai Le Quy	IU International University of Applied Sciences, Germany
Seiyon Lee	University of Florida, USA
Victor Lee	Stanford University, USA
Marie Lefevre	LIRIS - Université Lyon 1, France
Blair Lehman	Educational Testing Service, USA
Juho Leinonen	Aalto University, Finland
Liang-Yi Li	National Taiwan Normal University, Taiwan
Zhi Li	University of California, Berkeley, USA
Bibeg Limbu	University of Duisburg-Essen, Germany
Carla Limongelli	Università Roma Tre, Italy
Jionghao Lin	Carnegie Mellon University, USA
Allison Littlejohn	University College London, UK
Qi Liu	University of Science and Technology of China, China
Ziyuan Liu	Singapore Management University, USA
Sonsoles López-Pernas	University of Eastern Finland, Finland
Yu Lu	Beijing Normal University, China
Vanda Luengo	Sorbonne Université - LIP6, France
Collin Lynch	North Carolina State University, USA
John Magee	Clark University, USA
George Magoulas	Birkbeck, University of London, UK
Linda Mannila	Åbo Akademi University, Finland

Mirko Marras	University of Cagliari, Italy
Alejandra Martínez-Monés	Universidad de Valladolid, Spain
Goran Martinovic	J.J. Strossmayer University of Osijek, Croatia
Jeffrey Matayoshi	McGraw Hill ALEKS, USA
Diego Matos	Federal University of Alagoas, Brazil
Noboru Matsuda	North Carolina State University, USA
Guilherme Medeiros Machado	ECE Paris, France
Christoph Meinel	Hasso Plattner Institute, Germany
Roberto Angel Melendez Armenta	Universidad Veracruzana, Mexico
Agathe Merceron	Berlin State University of Applied Sciences, Germany
Eva Millan	Universidad de Málaga, Spain
Marcelo Milrad	Linnaeus University, Sweden
Wookhee Min	North Carolina State University, USA
Tsunenori Mine	Kyushu University, Japan
Sein Minn	Inria, France
Péricles Miranda	UFPE, Brazil
Tsegaye Misikir Tashu	University of Groningen, Netherlands
Tanja Mitrovic	University of Canterbury, New Zealand
Phaedra Mohammed	University of the West Indies, Trinidad and Tobago
Mukesh Mohania	IIIT Delhi, India
Bradford Mott	North Carolina State University, USA
Ana Mouta	USAL, Portugal
Chrystalla Mouza	University of Illinois at Urbana-Champaign, USA
Pedro J. Muñoz-Merino	Universidad Carlos III de Madrid, Spain
Takashi Nagai	Institute of Technologists, Japan
Tricia Ngoon	Carnegie Mellon University, USA
Huy Nguyen	Carnegie Mellon University, USA
Narges Norouzi	UC Berkeley, USA
Dr. Nasheen Nur	Florida Institute of Technology, USA
Benjamin Nye	University of Southern California, USA
Xavier Ochoa	New York University, USA
Jaclyn Ocumpaugh	University of Pennsylvania, USA
Elaine H. T. Oliveira	Universidade Federal do Amazonas, Brazil
Jennifer Olsen	University of San Diego, USA
Ranilson Paiva	Universidade Federal de Alagoas, Brazil
Viktoria Pammer-Schindler	Graz University of Technology, Austria
Luc Paquette	University of Illinois at Urbana-Champaign, USA
Abelardo Pardo	University of Adelaide, Australia
Zach Pardos	University of California, Berkeley, USA
Rebecca Passonneau	Pennsylvania State University, USA

Rumana Pathan	Indian Institute of Technology Bombay, India
Mar Perez-Sanagustin	Université Paul Sabatier Toulouse III, France
Yang Pian	Beijing Normal University, China
Niels Pinkwart	Humboldt-Universität zu Berlin, Germany
Eduard Pogorskiy	Open Files Ltd., UK
Oleksandra Poquet	Technical University of Munich, Germany
Ethan Prihar	École polytechnique fédérale de Lausanne, Switzerland
David Pynadath	University of Southern California, USA
Yuvaraj R.	Anna University, India
Mladen Rakovic	Monash University, Australia
Sowmya Ramachandran	Stottler Henke Associates Inc., USA
Steven Ritter	Carnegie Learning, Inc., USA
Luiz Rodrigues	Federal University of Alagoas, Brazil
José Raúl Romero	University of Cordoba, Spain
Carolyn Rose	Carnegie Mellon University, USA
Rinat B. Rosenberg-Kima	Technion - Israel Institute of Technology, Israel
José A. Ruipérez Valiente	University of Murcia, Spain
Sreecharan Sankaranarayanan	Carnegie Mellon University, USA
Eileen Scanlon	Open University, UK
Robin Schmucker	Carnegie Mellon University, USA
Bertrand Schneider	Harvard University, USA
Kazuhisa Seta	Osaka Metropolitan University, Japan
Lele Sha	Monash University, Australia
Tasmia Shahriar	North Carolina State University, USA
Kshitij Sharma	Norwegian University of Science and Technology, Norway
Bruce Sherin	Northwestern University, USA
Atsushi Shimada	Kyushu University, Japan
Valerie Shute	FSU, USA
Aditi Singh	Cleveland State University, USA
Sean Siqueira	Federal University of the State of Rio de Janeiro (UNIRIO), Brazil
Andy Smith	North Carolina State University, USA
Nancy Songer	University of Utah, USA
John Stamper	Carnegie Mellon University, USA
Christian M. Stracke	University of Bonn, Germany
Abhijit Suresh	University of Colorado Boulder, USA
Vinitra Swamy	EPFL, Switzerland
Paula Toledo Palomino	Sao Paulo State College of Technology (FATEC) - Matão, Brazil
Maya Usher	Technion, Israel

Masaki Uto	University of Electro-Communications, Japan
Lisa van der Heyden	University of Duisburg-Essen, Germany
Kurt Vanlehn	Arizona State University, USA
Julita Vassileva	University of Saskatchewan, Canada
Esteban Vázquez-Cano	UNED, Spain
Olga Viberg	KTH Royal Institute of Technology, Sweden
Francisco Vicente-Castro	New York University, USA
Esteban Villalobos	Université Paul Sabatier, France
Alessandro Vivas	UFVJM, Brazil
Erin Walker	Arizona State University, USA
Ning Wang	University of Southern California, USA
Christabel Wayllace	New Mexico State University, USA
Stephan Weibelzahl	Private University of Applied Sciences Göttingen, Germany
Gary Weiss	Fordham University, USA
Chris Wong	University of Technology Sydney, Australia
Beverly Park Woolf	University of Massachusetts, USA
Qian Xiao	Trinity College Dublin, Ireland
Wanli Xing	University of Florida, USA
Masanori Yamada	Kyushu University, Japan
Sho Yamamoto	Kindai University, Japan
Zhang Yupei	Northwestern Polytechnical University, China
Andrew Zamecnik	University of South Australia, Australia
Jianwei Zhang	University at Albany, USA
Shu-Gang Zhang	Ocean University of China, China
Yong Zheng	Illinois Institute of Technology, USA
Gaoxia Zhu	National Institute of Education, Nanyang Technological University, Singapore
Stefano Zingaro	Università di Bologna, Italy
刘强	中国四川成都树德中学

International Artificial Intelligence in Education Society

Management Board

President

Olga C. Santos	UNED, Spain

Secretary/Treasurer

Vania Dimitrova	University of Leeds, UK

Journal Editors

Vincent Aleven	Carnegie Mellon University, USA
Judy Kay	University of Sydney, Australia

Finance Chair

Benedict du Boulay	University of Sussex, UK

Membership Chair

Benjamin D. Nye	University of Southern California, USA

IAIED Officers

Yancy Vance Paredes	North Carolina State University, USA
Son T. H. Pham	Nha Viet Institute, USA
Miguel Portaz	UNED, Spain

Executive Committee

Akihiro Kashihara	University of Electro-Communications, Japan
Amruth Kumar	Ramapo College of New Jersey, USA
Christothea Herodotou	Open University, UK
Jeanine A. Defalco	CCDC-STTC, USA
Judith Masthoff	Utrecht University, Netherlands
Maria Mercedes T. Rodrigo	Ateneo de Manila University, Philippines
Ning Wang	University of Southern California, USA
Olga Santos	UNED, Spain
Rawad Hammad	University of East London, UK
Zitao Liu	Jinan University, China
Bruce M. McLaren	Carnegie Mellon University, USA
Cristina Conati	University of British Columbia, Canada
Diego Zapata-Rivera	Educational Testing Service, Princeton, USA
Erin Walker	University of Pittsburgh, USA
Seiji Isotani	University of São Paulo, Brazil
Tanja Mitrovic	University of Canterbury, New Zealand

Noboru Matsuda	North Carolina State University, USA
Min Chi	North Carolina State University, USA
Alexandra I. Cristea	Durham University, UK
Neil Heffernan	Worcester Polytechnic Institute, USA
Andrew M. Olney	University of Memphis, USA
Irene-Angelica Chounta	University of Duisburg-Essen, Germany
Beverly Park Woolf	University of Massachusetts, USA
Ig Ibert Bittencourt	Universidade Federal de Alagoas, Brazil

Anticipating Student Abandonment and Failure: Predictive Models in High School Settings

Emanuel Marques Queiroga[1,2,3](✉), Daniel Santana[4], Marcelo da Silva[4], Martim de Aguiar[2], Vinicius dos Santos[4], Rafael Ferreira Mello[5], Ig Ibert Bittencourt[3], and Cristian Cechinel[3,5,6]

[1] Federal Institute of Education, Science and Technology Sul-rio-grandense (IFSul), Pelotas, Brazil
emanuelmqueiroga@gmail.com
[2] Inter American Development Bank (IDB), Brasília, Brazil
[3] Center of Excellence for Social Technologies (NEES), Federal University of Alagoas (UFAL), Maceió, Brazil
ig.ibert@ic.ufal.br, contato@cristiancechinel.pro.br
[4] Instituto Unibanco, São Paulo, Brazil
{daniel.santos-santana,marcelo.pessoa-silva, vinicius.oliveira-santos}@sgpiu.org.br
[5] Centro de Estudos e Sistemas Avançados do Recife (CESAR), Recife, Brazil
rafael.mello@ufrpe.br
[6] Federal University of Santa Catarina (UFSC), Florianópolis, Brazil

Abstract. Addressing the issue of high school non-completion poses a crucial challenge for contemporary education. This research introduces a machine learning-based methodology to identify students at risk of failure and abandonment in a specific Brazilian state, aiming to establish an early warning system utilizing academic, socioeconomic, and performance indicators for proactive interventions. The methodology followed here ensures the explainability of predictions and guards against bias in relation to certain features. The analysis of data from 79,165 students resulted in the creation of six accurate classification models, with accuracy rates ranging from 69.4% to 92.7%. This underscores the methodology's effectiveness in identifying at-risk students, highlighting its potential to alleviate failure and abandonment. The implementation of this methodology could positively influence proactive educational policies and enhance educational metrics within the state.

Keywords: Learning Analytics · Educational Data Mining · Educational Technology

1 Introduction

High school education in Brazil is characterized by a complex interaction among the education systems of states and the Union, resulting in a diverse, decentralized, and challenging landscape. Each federative entity has the autonomy

A. M. Olney et al. (Eds.): AIED 2024, LNAI 14829, pp. 351–364, 2024.
https://doi.org/10.1007/978-3-031-64302-6_25

to establish its own educational policies, leading to significant variations in the quality of education provided [26]. This decentralization is reflected in socioeconomic disparities, school infrastructure conditions, and student performance indices by state, contributing to the creation of a complex and multifaceted educational scenario [17,22,26].

Despite clear efforts made, Brazilian Education faces significant challenges, with dropout, abandonment and failures as the most prominent issues. School dropout is characterized by the student leaving school before completing his/her studies, whether abandonment is when the student drops the school in a given year but returns in the next. Failure consists the repetition of academic year [12]. All these issues compromise the access to education, the quality of learning, and the consequent development of the students.

The performance of students in school is normally compromised by multifactorial problems, making the understanding and prevention of the risks of academic failure complex [1]. Thus, in search of alternatives and mechanisms to assist educational stakeholders and schools in developing and monitoring evidence-based public policies, the use of predictive analysis emerges as one of the main tools for obtaining actionable insights from available data [2]. The use of such an approach helps educational stakeholders identify students' behavioral patterns, providing a more robust understanding of the factors contributing to dropout and failure. Predictive analysis has the potential to early identify students at risk of successfully completing their academic trajectory, enabling the implementation of preventive policies for them [10,15].

The current paper outlines the development of an early warning system designed to predict students at risk of failure and/or abandonment in the State of Espírito Santo, Brazil. The proposed methodology integrates various sets of features, including school variables and socioeconomic information, as input for generating machine learning models to predict students at risk at an early stage. Furthermore, the proposed methodology conducts a thorough bias analysis, aiming for models that are equitable and fairness. The results obtained align with the existing literature, offering a new perspective on implementation strategies for educational intervention in real life.

2 Related Literature

The field of educational data analysis has evolved significantly, with various studies demonstrating the effectiveness of applying Artificial Intelligence algorithms in identifying students at risk in high school [9]. However, these studies diverge in the data used and the contexts of application. In addition to socioeconomic factors, prior academic performance emerges as a powerful risk indicator [14,17,20]. Students with a history of low performance in fundamental subjects are more likely to fail [14,20]. Early identification of difficulties and the implementation of targeted interventions are crucial to prevent further academic delays [25].

A multifactorial approach is crucial to understanding school abandonment and academic failure [13]. This approach may involve a diverse range of factors that impact the student, including significant socioeconomic aspects [14]. Studies have consistently shown that students from low-income families face greater academic and persistence challenges [23]. The lack of educational resources, limited family support, and socioeconomic adversities contribute to decreased engagement and academic success [21,29].

The analysis of these features should not occur in isolation but rather as part of a holistic system that integrates socioeconomic, emotional, and academic factors. This approach provides a more comprehensive view of individual student circumstances and allows for the development of more personalized support strategies [8,28]. Early identification, based on a variety of indicators, enables more effective interventions at critical moments in the student's educational journey, reducing the risks of dropout, abandonment and academic failure.

In this way, the studies by Márquez-Vera et al. [20], Parr & Bonitz [23], Chung and Lee [5], Gómez et al. [13], and Krüger et al. [17] collectively offer a diverse perspective on the use of data analytics and machine learning in predicting school dropout rates. At the same time, the search for educational patterns that can assist in identifying students with high probabilities of failure is a general concern in the field, with studies such as that of Hernandez-Leal et al. [14] showing that in the transition from elementary to secondary education, difficulties in Mathematics and Language subjects are critical points, significantly impacting student failure rates, regardless of socioeconomic level or school location.

Márquez-Vera et al. [20] and Chung and Lee [5] both implemented machine learning techniques - the former in Mexico and the latter in Korea. Márquez-Vera et al. achieved an impressive 99.8% accuracy in predicting early school dropout, demonstrating the power of data mining in educational settings. Similarly, Chung and Lee's [5] use of random forests in a large dataset of Korean students showed a 95% accuracy rate, underscoring the efficacy of machine learning algorithms in identifying at-risk students.

On the other hand, Parr & Bonitz [23] in the USA and Gómez et al. [13] in Chile approached the issue through different lenses. Parr & Bonitz's study focused on the influence of family background and student behavior, emphasizing socio-economic and behavioral factors, while Gómez et al. [13] developed models for Chile's education system, highlighting the need for large-scale data analysis in educational policy and intervention planning.

In the final analysis, the study by Krüger et al. [17] in Brazil stands out for its use of an explainable machine learning approach, achieving a notable accuracy of up to 95%. This research highlights the critical aspect of model interpretability, ensuring that educators and policymakers can understand and utilize the insights provided by such predictive models. These diverse studies, summarized in Table 1, provide a comprehensive overview encompassing objectives, levels of application, data used, types of data, results obtained, and conclusions, showcasing the multifaceted nature of data analytics in educational contexts across different countries.

In this study, we propose a customized approach for the early identification of high school students at risk of abandonment or failure in Espírito Santo, Brazil.

Table 1. Comparative of State of Art

Study	Goal	Educational Level	Data Used	Types of Data	Results	Conclusion
Márquez-Vera et al. [20]	Predict early school dropout	High School in Mexico	419	Student academic and personal data	99.8% ACC	Effectiveness of data mining in dropout prediction
Parr & Bonitz [23]	Influence of family background on school dropout	High School in the USA	15,753	Demographic, behavioral, academic performance data	-	Importance of socio-economic and behavioral factors
Chung & Lee [5]	Predict school dropout using ML	High School in Korea	165,715	Comprehensive educational data	95% ACC	Effectiveness of random forests in risk identification
Gómez et al. [13]	Develop ML models	High School in Chile	Large datasets	Administrative educational data	-	Need for large-scale approaches
Krüger et al. [17]	Explainable machine learning for dropout prediction	High School in Brazil	19 schools Data	School records, academic performance	95% ACC	Importance of interpretable models in education
Hernandez Leal et al. [14]	Unveil educational patterns in Colombia	High School in north of Colombia	6,400	Academic performance, socio-economic data	60% - 99,8% F1	Importance of identifying educational patterns for policy- making

This approach differs from the literature in some main aspects. The first is the large-scale perspective and development through the needs of the State's Department of Education. The second is that it is developed with practical integration in mind, which will be implemented in the next stage of the project. Third, this approach significantly differs from the methodologies found in the reviewed literature by combining academic, socioeconomic, and performance data to create a predictive model adapted to local specifics. At last, another important aspect is the use of fairness analysis to ensure algorithmic equity.

3 Context of the Experiments

The state of Espírito Santo, located in the southeastern region of Brazil, stands out for its unique educational characteristics. In 2021, the Basic Education Development Index (IDEB) in the state was 4.4 [16], while the established goal was 4.9, highlighting the need for improvements to achieve the proposed objectives. Additionally, the age-grade distortion reached 20.9%, rising to 25.5% in the first year. These numbers indicate significant challenges in the proper progression of students throughout the school years.

In 2022, the state's secondary education system in Espírito Santo faced a 6.6% repetition rate and a 2% abandonment rate, accounting for 8,884 students who encountered challenges in their education trajectory [16]. Although these indicators represent an improvement compared to the historical school performance results, obstacles in promoting continuous academic progression still persist. In comparison, Brazil recorded rates of 7.7% and 5.7% for repetition and abandonment, respectively [16].

Despite better indicators than the national average, critically assessing educational challenges in Espírito Santo is crucial. Understanding factors behind abandonment and failure is key for effective preventive measures. Continuous monitoring of educational indicators should guide policy development for improving education quality and efficiency in the state.

4 Methodology

This study applies the Cross-Industry Standard Process for Data Mining (CRISP-DM) model to guide the analysis of educational data [27]. CRISP-DM is a robust model that facilitates understanding, preparation, modeling, evaluation, and deployment of data in a structured process.

4.1 Data Collection and Processing

Data collection adhered to Brazil's General Data Protection Law (LGPD), ensuring student data privacy and security [19]. Similar to the European Union's GDPR, LGPD underscores transparency, consent, and personal data protection [19]. The data were anonymized to comply with these regulations.

In this stage, data manipulation was performed to structure the available information appropriately. Data cleaning procedures, treatment of missing and duplicate information, outlier detection, and selection of the population and features of interest were carried out. The standardization and structuring strategy followed the steps:

1. **Selection of the reference**: High school students from Espírito Santo with valid enrollment.
2. **Selection of the features**: It was chosen to exclude columns with constant and null values, duplicates and other information that did not apply to the study.
3. **Handling of duplicate observations**: We applied the split-apply-combine technique so that each row in the database be a unique record.
4. **Missing data treatment**: We applied proxy technique to handle information that existed but was not recorded.
5. **Exclusion of observations**: Exclusion of duplicate students IDs preserving the most recent enrollment date.

4.2 Data Description

Table 2 presents the number of students in the dataset. These data encompass information from the academic year 2022, including a total of 79,165 students enrolled in high schools in the state of Espírito Santo, Brazil. The database represents a diverse mix of institutions in terms of geographical location (urban and rural), covering all 78 municipalities in the state.

Table 2. Total amount of students

School Year	Total of Students	Approved	Failed and/or abandoned
First Year	29,101	24,705	4,396
Second Year	27,017	23,656	3,361
Third Year	23,047	21,711	1,336
Total	79,165	70,072	9,093

The student demographic is diverse, encompassing a wide age range typical of high school, which in Brazil is from 14 to 17 years old. However, these values show a slight tendency to increase due to age-grade distortion. The gender distribution is approximately balanced, with 45% boys and 55% girls. Regarding the ethnic composition of students, it reflects the cultural diversity of the state. Additionally, a total of 39% of students are enrolled in government assistance programs, which is a significant percentage, indicating a considerable presence of students from low socioeconomic backgrounds.

Groups of Features. The features used to generate the predictive models can be broadly classified into three different groups in accordance with their educational importance, as follows:

- **Demographic and socioeconomic**: The features of age, gender, ethnicity, and social government assistance status are essential for understanding students' context. Literature suggests these socioeconomic factors significantly influence access to educational resources and extracurricular support, impacting student performance and motivation [21,29].
- **Academic**: features such as student grades, absences, and attendance are employed, providing a direct measure of student performance and engagement. Low performance and a high rate of absences are associated with a greater risk of dropout and failure, requiring specific educational interventions [24,25].
- **Disability-related issues**: features such as the presence of any type of disability and the need for any kind of special assistance are employed. Including information about the type of disability and support resources ensures an inclusive educational approach. Adaptive educational resources for students with disabilities are important to minimize negative impacts on academic performance [3,11].

Fairness Analysis. The features were examined to identify any potential for generating unwanted biases in the automated models. The exploratory data analysis included an assessment not only of the significance of each attribute but also of how their values are distributed. As a result, five features were identified for in-depth analysis as they could be potential generators of discriminatory biases: "*Ethnicity*", "*Government Assistance*," "*Gender*," "*Class Period*," and "*Disability*". These five features were selected after several meetings with the Department of Education of the State of Espírito Santo and based on the existing related literature about bias in predictive learning analytics [7,29].

4.3 Predictive Models

The predictive models used in this study are *white box*, emphasizing the commitment to transparency in algorithmic decision-making processes. These models are known for their interpretability, providing clear insights into how input features influence predictions [17]. In educational contexts, where decisions shape students' paths, interpreting predictions is vital to foster understanding among educators and administrators, to build trust in the analytical tools, and to give transparency for auditing and review [15].

The following *white-box* algorithms were tested: Logistic Regression, AdaBoost Classifier, RandomForest Classifier, DecisionTree Classifier, MLP Classifier (used as a benchmark), Gaussian Naive Bayes, and HistGradient Boosting Classifier. The inclusion of MLP is to provide a comparative performance of more complex models [13]. The models were trained and tested using 10-fold cross-validation and evaluated using the following metrics: accuracy, precision, recall, AUCROC, confusion matrix, and f1-measure.

Data was balanced with random oversampling and undersampling techniques and using the imblearn library [18]. This strategy balanced the classes distribution by increasing the representation of minority classes (students that failed and/or abandoned) through oversampling and reducing the dominance of majority classes through undersampling (students that approved).

The modeling strategy involved using two models for each high school year, one at the start of the academic year and another at the end of the first trimester. This dual-phase approach is essential for capturing the dynamic nature of the educational environment and students' progress. The choice of this two-stage strategy is primarily grounded in the following factors:

- **Temporal dynamics of student development**: The performance and needs of students evolve throughout the academic year. Models generated at different times can capture these changes, enabling more precise and timely interventions [4,5,21].
- **Adaptation to changes in the school environment**: New challenges and opportunities arise during the academic year. A model generated after the first trimester can incorporate recent data, better reflecting the current situation of students [10,15].

– **Early prediction and continuous monitoring**: An initial model provides the ability to early identify students at risk, while the first-trimester model allows for adjusting and refining these predictions as more information becomes available.

Figure 1 depicts the 6 models developed according to the mentioned strategy (2 for each year grade of high school).

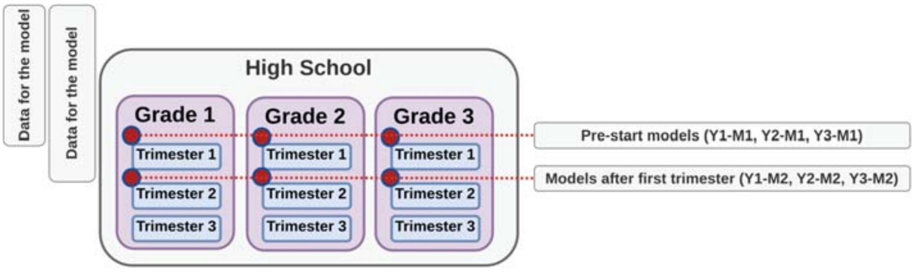

Fig. 1. Predictive models developed for the three years of high school

Depending on the moment of the school year the model is used, different input information is provided, as follows:

1. **Pre-start Models**: Based on historical data, this model utilizes features such as demographic and socioeconomic data to predict risks before the start of the academic year. These models do not use previous academic performance features as they were not available in the dataset.
2. **Models after first trimester**: Built after the end of the first trimester, this model incorporates recent data, including performance in the current trimester and other relevant features such as demographic data and socioeconomic data to refine the initial predictions. These models also use performance-derived features, such as the student's position relative to the class average, to place the student within the school context.

5 Results

5.1 Performance of the Predictive Models

Table 3 displays the performance of the best predictive models for each stage of students' high school trajectory. In the $Model$ column, Y represents the school year, and M indicates the moment of model application. As shown in the table, models $Y1-M2$, $Y2-M2$, and $Y3-M2$ exhibit the best performances, with high accuracies, indicating a consistent and accurate identification of students at risk of failing or abandon the school year. These results also indicate that as more information is used in the models, the better will be their performances during the year. An example of this situation is depicted in the feature selection of the top twenty features ($F = 20$) of the $Y1-M2$ model, as displayed in Fig. 2, Notably, the influence of features representing grades, attendance, and absences throughout the first trimester of the first high school year is evident

Table 3. Results of best classifier for each model

Model	Classifier	Sampling	Accuracy	Precision	Recall	AUC	F1	Confusion Matrix (TP, FN FP, TN)
Y1-M1	Random Forest	Feature Selection (F=20) Random Undersampling	69.3	28.46	68.99	68.78	40.24	1289.1, 565.9 102.3, 225.2
Y1-M2	Random Forest	Feature Selection (F=20) Random Undersampling	90	81.54	50.43	74.2	62.29	1557.6, 293.7 44.3, 286.9
Y2-M1	Random Forest	Random Undersampling	69.4	23.83	67.72	68.67	35.24	1237.2, 539.7 80.3, 169.0
Y2-M2	Random Forest	Random Undersampling	92.7	85.41	51.45	75.08	64.12	1512.0, 258.5 36.4, 219.3
Y3-M1	Random Forest	Random Undersampling	82.54	24.51	75.21	79.11	33.31	1351.2, 276.6 25.1, 75.6
Y3-M2	Random Forest	Random Undersampling	89.6	34.7	86.95	88.35	49.51	1460.3, 166.6 13.1, 88.5

As it is possible to see, the best models were all generated by the Random Forest algorithm. Analyzing specific metrics, we note that models $Y2-M2$ and $Y3$–$M2$ have excellent areas under the ROC curve (AUC), indicating remarkable discrimination. This suggests that these models are effective in distinguishing students with and without academic issues. The balanced F1-Score of model $Y2-M2$ highlights its good balance between precision and recall, signaling a harmonious classification ability, especially at the end of the first trimester of the academic year.

An important analysis focuses on the true negatives (TN) present in the confusion matrices of Table 3, representing the correct classification of the students who will fail or abandon. These results play a significant role, particularly when considering the remarkable stability at the end of the first academic trimester (M2).

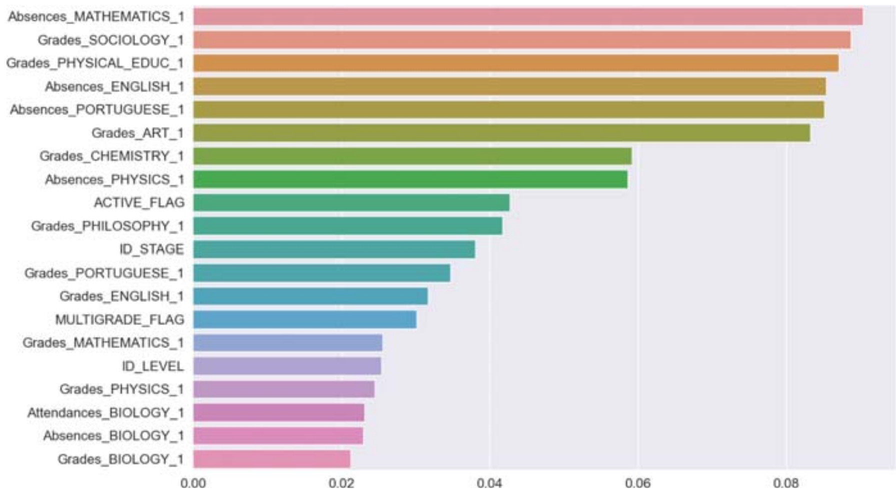

Fig. 2. Feature Selection (F = 20 for Y1-M2.

The results suggest that the models, notably those at the end of the first trimester, have promising potential for early identification of students with possible academic trajectory issues. These findings have valuable implications for intervention strategies and personalized support, aiming to enhance students' academic paths throughout the academic period.

5.2 Fairness Analysis

The bias of the models developed in this study were assessed using the What-If Tool. This tool allows the evaluation of how model decisions may vary for different groups or student characteristics. Bias analysis is crucial in the educational context, where equity and justice in education are fundamental priorities. Ensuring that prediction models do not perpetuate or amplify existing inequalities is a crucial objective, and the What-If Tool provides a systematic and transparent approach to achieving this goal.

Table 4 presents the results of the bias analysis for each models according to the protected features.

As can be seen in the Table 4, the accuracy values among the classes of features with potential for generating undesirable bias show that none of the models exhibited bias concerning the features defined as potential sources of bias. This positive result highlights the effectiveness of the strategies adopted to mitigate any unfair tendencies in the prediction models.

Table 4. Model accuracy by class for features with potential bias, illustrating bias mitigation effectiveness.

Model	Gender		Class Period		Government Assistance		Ethnicity		Disability	
	Male	Female	Day	Evening	Yes	No	White	Non White	Yes	No
Y1-M1	79.1	78.9	81.8	78.7	77.8	80.2	78.8	85	71.4	79.3
Y1-M2	87.7	83.3	85	90	85.4	85.4	88.5	83.1	85.3	87.5
Y2-M1	71.5	75	88.6	71.7	72.6	73.6	73.3	71.4	87.5	72.5
Y2-M2	87	87.4	86	98	86.9	87.4	87	92.9	81	87,5
Y3-M1	65.4	73.9	70	69.6	65.3	73	63	79.4	80	69.8
Y3-M1	91.6	92	91	98.1	91.5	92	93.8	90.7	81.3	92.1

6 Discussion

This work presents a methodology for creating prediction models for early risk of abandonment or failure in high school, demonstrating the efficacy and applicability of machine learning techniques in identifying patterns and predictors for that purpose. The research aligns with the existing literature in learning analytics, a subfield of artificial intelligence and machine learning, highlighting the accuracy of these techniques to enhance educational interventions and promote student success [21].

Compared to previous works, such as those by Márquez-Vera et al. [20] and Chung and Lee [5], this study distinguishes itself by integrating academic, socioeconomic, and performance data, adopting a multifactorial approach, and emphasizing the interpretability of models, in line with current trends in explainable machine learning, as proposed in the work of Kruger et al. [17]. Moreover, the research stands out for its focus on eliminating biases of algorithmic discrimination, an aspect not yet significantly addressed in the field's literature, promoting fair and equitable analytical models, reflecting an advancement in the theory on the subject in Learning Analytics, as proposed by Deho et al. [7].

The work also distinguishes itself in terms of its scope, and applicability on a large scale, focusing on a significant number of students. Compared to the works of Márquez-Vera et al. [20], Parr & Bonitz [23], Krüger et al. [17], and Hernandez-Leal et al. [14], the student numbers are significantly larger, covering the entire public high school network of the state. Thus, unlike previous research that may have been limited to smaller contexts or specific samples [14,17,20,23], this study demonstrates robustness and flexibility that allow its application in different educational contexts, covering a significant number of students. This scalability ensures that the insights and interventions proposed can be adapted and implemented in educational systems of various sizes and characteristics in other Brazilian states, provided small adjustments are made to the educational needs and contexts, offering a viable solution for large-scale educational improvements.

The participatory approach in methodological development, closely collaborating with the Secretary of Education of Espírito Santo (SEDU-ES), is a relevant aspect of the practical implementation of the work, breaking down the last barrier established by Clow et al. [6] in the practical application of LA. Furthermore, this approach ensures that the proposed educational strategies are effective and adapted to local specificities, offering a viable and more adaptable model to the context and user expectations, thus achieving higher levels of acceptance as established by the research of Herodotou et al. [15].

The research underlines the importance of using data to improve teaching and learning processes, highlighting the need for personalized educational strategies and interventions that address socioeconomic disparities [3,11]. By providing detailed data and analyses, the study offers a basis for the formulation of affirmative policies that promote equity and inclusion, contributing to sustainable economic development through the improvement of educational indices, seeking to ensure fair opportunities for all students.

7 Final Remarks

This study demonstrated the feasibility of using machine learning techniques for early identification of high school students at risk in their academic journey in Espírito Santo. The developed models, applying academic, socioeconomic, and performance data, showed high accuracy, highlighting their potential in early prediction of school abandonment and failure.

The research emphasizes the importance of analytical tools in education, especially in regions with significant challenges in school failure. The study offers valuable insights for proposing a data-driven approach to mitigate school dropout.

The proposed models can serve as crucial tools for educators and school administrators, allowing for targeted and timely interventions. They represent a significant step towards a more adaptive and responsive educational system to student needs.

Although the results are promising, the study faces limitations, such as the need to evaluate the acceptance of the models by educators. Furthermore, future research could explore the inclusion of more features and the application of the models in other regions to generalize the findings. An important step for improving accuracy rates, especially in models at the beginning of the school year, is the inclusion of students' historical data, as proposed by Queiroga et al [25].

This study underscores the relevance of data-based approaches in education, offering a new perspective in the fight against school abandonment and failure. It sets a precedent for the implementation of innovative and effective educational strategies, highlighting the potential of analytical technologies in improving educational quality.

Acknowledgments. This work was supported by Inter American Development Bank and Unibanco Institute. Cristian Cechinel was partially supported by the Brazilian National Council for Scientific and Technological Development (CNPq) (grants 305731/2021-1 and 409633/2022-4).

References

1. Amparo, D.M.d., Galvão, A.C.T., Cardenas, C., Koller, S.H.: A escola e as perspectivas educacionais de jovens em situação de risco. Psicologia Escolar e Educacional **12**(1), 69-88 (2008). https://doi.org/10.1590/S1413-85572008000100006
2. Attaran, M., Attaran, S.: Opportunities and challenges of implementing predictive analytics for competitive advantage. In: Applying Business Intelligence Initiatives in Healthcare and Organizational Settings, pp. 64–90 (2019)
3. Cechinel, C., et al.: Mapping learning analytics initiatives in Latin America. Br. J. Educ. Technol. **51**(4), 892–914 (2020)
4. Chatti, M.A., Dyckhoff, A.L., Schroeder, U., Thüs, H.: A reference model for learning analytics. Int. J. Technol. Enhanced Learn. **4**(5–6), 318–331 (2013)
5. Chung, J.Y., Lee, S.: Dropout early warning systems for high school students using machine learning. Child Youth Serv. Rev. **96**, 346–353 (2019)
6. Clow, D.: The learning analytics cycle: closing the loop effectively (2012)
7. Deho, O.B., Zhan, C., Li, J., Liu, J., Liu, L., Le Duy, T.: How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics? Br. J. Edu. Technol. **53**(4), 822–843 (2022)
8. Elistia, E., Syahzuni, B.A.: The correlation of the human development index (HDI) towards economic growth (GDP per capita) in 10 Asean member countries. JHSS (J. Hum. Soc. Stud.) **2**(2), 40–46 (2018)
9. Esbenshade, L., Vitale, J., Baker, R.S.: Non-overlapping leave future out validation (nolfo): Implications for graduation prediction. prePrint . https://doi.org/10.35542/osf.io/tx8cy
10. Ferguson, R.: Learning analytics: drivers, developments and challenges. Int. J. Technol. Enhanced Learn. **4**(5/6), 304–317 (2012)
11. Ferguson, R., et al.: Research evidence on the use of learning analytics: implications for education policy (2016)
12. Ferreira, S.G., Ribeiro, G., Tafner, P.: School abandonment and dropout in Brazil. Technical report, Rio de Janeiro, Brazil (2022)
13. Gómez-Pulido, J.A., Park, Y., Soto, R.: Advanced techniques in the analysis and prediction of students' behaviour in technology-enhanced learning contexts (2020)
14. Hernández-Leal, E., Duque-Méndez, N.D., Cechinel, C.: Unveiling educational patterns at a regional level in Colombia: data from elementary and public high school institutions. Heliyon **7**(9), e08017 (2021)
15. Herodotou, C., Rienties, B., Boroowa, A., Zdrahal, Z., Hlosta, M., Naydenova, G.: Implementing predictive learning analytics on a large scale: the teacher's perspective. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference, pp. 267–271 (2017)
16. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP): Censo Escolar 2021 (2022). https://qedu.org.br/uf/32-espirito-santo. Accessed 1 Jan 2024
17. Krüger, J.G.C., de Souza Britto Jr., A., Barddal, J.P.: An explainable machine learning approach for student dropout prediction. Expert Syst. Appl. **233**, 120933 (2023)

18. Lemaître, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Rese. **18**(17), 1–5 (2017). http://jmlr.org/papers/v18/16-365.html
19. Lorenzon, L.N.: Análise comparada entre regulamentações de dados pessoais no brasil e na união europeia (lgpd e gdpr) e seus respectivos instrumentos de enforcement. Revista do Programa de Direito da União Europeia **1**, 39–52 (2021)
20. Márquez-Vera, C., Cano, A., Romero, C., Noaman, A.Y.M., Mousa Fardoun, H., Ventura, S.: Early dropout prediction using data mining: a case study with high school students. Expert Syst. **33**(1), 107–124 (2016)
21. Moissa, B., Gasparini, I., Kemczinski, A.: A systematic mapping on the learning analytics field and its analysis in the massive open online courses context. Int. J. Distance Educ. Technol. (IJDET) **13**(3), 1–24 (2015)
22. OECD: Benchmarking Higher Education System Performance. OECD Publishing, Paris, France (2019). https://doi.org/10.1787/be5514d7-en
23. Parr, A.K., Bonitz, V.S.: Role of family background, student behaviors, and school-related beliefs in predicting high school dropout. J. Educ. Res. **108**(6), 504–514 (2015)
24. Queiroga, E.M., et al.: Experimenting learning analytics and educational data mining in different educational contexts and levels. In: 2022 XVII Latin American Conference on Learning Technologies (LACLO), pp. 1–9. IEEE (2022)
25. Queiroga, E.M., Batista Machado, M.F., Paragarino, V.R., Primo, T.T., Cechinel, C.: Early prediction of at-risk students in secondary education: A countrywide k-12 learning analytics initiative in Uruguay. Information **13**(9), 401 (2022)
26. Stanek, C.: The educational system of Brazil. IEM Spotlight **10**(1), 1–6 (2013)
27. Wirth, R., Hipp, J.: Crisp-dm: Towards a standard process model for data mining. In: Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, vol. 1, pp. 29–39. Springer, London (2000)
28. Yakunina, R., Bychkov, G.: Correlation analysis of the components of the human development index across countries. Procedia Econ. Finan. **24**, 766–771 (2015)
29. Yau, J.Y.K., Ifenthaler, D.: Reflections on different learning analytics indicators for supporting study success. Int. J. Learn. Anal. Artif. Intell. Educ.: IJAI **2**(2), 4–23 (2020)