Journal of Higher Education and Learning
ISSN 2998-7679
Vol. 2, No. 3, 2025
Published by STSL Press

Original Paper

Integrating Data Science into Secondary STEM Curriculum

William H. Robertson

Professor of STEM Education, Teacher Education Department, College of Education, The University of Texas at El Paso, Texas, USA

E-mail: robertson@utep.edu; (915) 747-6426

Abstract

The integration of data-driven approaches in education is also increasingly shaping secondary curricula, particularly in areas like funding forecasts, enrollment patterns, and career readiness. However, there is a concern that reliance on software tools, assumed to be free from bias, might influence educational outcomes in ways that are not fully understood. This is especially pertinent in the context of science, technology, engineering, and mathematics (STEM) education, where there is a pressing need to ensure that curricula remain both relevant and of high quality. The rapid growth of data science, reflected in technological advancements and the rising demand for computational skills, is influencing how secondary schools adapt to meet the demands of the digital age, reshaping how STEM subjects are taught and learned.

Keywords: Data science, STEM, curriculum, computer, machine learning

Introduction

In the current state of society, the use of technology that integrates large data sets is everywhere, and the devices we use daily, such as cell phones, computers, and tablets, integrate apps and functionality that produce outcomes that are seemingly individually directed. Yet, there may be an inherent bias within the decision-making process in any machine learning or artificial intelligence solution that integrates data, as the outcome is only as reliable as the data that goes into it. There is also a great need for quality and relevant education as it relates to science, technology, engineering and mathematics (STEM) classes and curriculum.

A basic decision-making process within a software solution involves collecting data, extracting patterns and facts from that data and using those patterns to make decisions. As data sets grow larger, there is a greater need to have variables, such as age, race or gender, become operationalized scores (Larson, et. al, 2016). In this way, the output is only as reliable as the data it analyzes, and without a mechanism to refine the data to include more current information, the ability of any machine learning or artificial intelligence platform is subject to inconsistencies. This is a fundamental in data science and one way that such study can be enhanced by researchers and teachers working in tandem. While the idea of data-driven solutions is quite popular in many phases of education, including projections for funding, enrollments, career pathways and retention, there is also an inherent reliability on the use of software solutions which are presumed to be without bias in determining outcomes.

In terms of teaching and learning, the fact that data science has experienced an exponential growth in recent time is reflected in the advancements and proliferation of technology, the increased demand for computing solutions and the proliferation of digital transformations. For teachers working with students, whether at the secondary or higher-education levels, this growth has also produced impacts in the job market, where the demand for software developers, data scientists, artificial engineers or cybersecurity experts has skyrocketed (The Importance of Computer Science Education, 2023). For students themselves, the knowledge of data science and its intricacies lead to impact their future and the ways in which they will live and work, and will impact important societal concerns, including healthcare accessibility, global climate change and resource availability.

Methodology and connection to data science

To provide opportunities for teachers and students to engage in real world research involving data science, cybersecurity and artificial intelligence, it is within this spirit that a collaborative effort from the Department of Computer Science and the Department of Teacher Education at The University of Texas at El Paso (UTEP) has begun to develop an immersive year-long professional development effort for secondary STEM teachers that focuses on Data Science, Artificial Intelligence and Cybersecurity. This effort will bring a local cohort of secondary STEM teachers to work in a six-week summer institute alongside researchers at UTEP to explore these issues and to develop a curriculum that can be integrated into their classrooms during the academic year.

The overarching goal of CREEDS Site is to provide an authentic research environment to STEM enthusiast Middle/High-School educators of the Paso Del Norte region alongside proficient UTEP Computer Science mentors. This co-learning environment will not only allow local educators to collaborate with UTEP researchers in addressing cyber-related research challenges via a project-based learning (PBL) approach but also empower their students through the delivery of curricular modules to potentially spark excitement for STEM among Middle / High-School (MHS) students. The exposure to cutting-edge, dynamic, and high-payoff CS areas, such as data science and cybersecurity, will have an early impact on MHS students, raising their awareness about STEM R&D and about the threats and risks of cyberspace.

The use of data science in the modern K-12 classroom has many implications and impacts. As a new technology, it can be something of high value in terms of developing critical thinking and conversely, it can be a subversive tool that can accelerate cheating and the development of misconceptions, especially in the areas of science, technology, engineering and mathematics (STEM). For many teachers, the concepts that originate in computer science and permeate the STEM fields in education, are rarely addressed in substantive manners in which their professional knowledge and skills are impacted positively. Occurring within this broader context, we will next describe a particular case of how data science is being integrated into a Computer Science Teacher Preparation Program. This example illustrates that we can begin to think about solving the more complex problems involved with integrating data science into any fields of education that are resistant, by first starting with those that are more receptive such as teachers of Computer Science.

Findings and Analysis

To facilitate an effective research environment and collaborative curricular design/delivery, the site will recruit pairs of teacher participants from the same school. In this way, the teachers would be comfortable in collaboratively designing the curricular component. Since the teacher participants are not required to have prior research experience, it the sole responsibility of the UTEP team to provide a platform where they can excel and successfully fulfill the mission of this project. To address the knowledge disparity among teacher participants, we provide significant background knowledge on different computer science topics in the first 2-weeks of the summer institute. In those tutorial sessions, the participants will gain first-hand knowledge along with hands-on experience on data science rich topics such as artificial intelligence, machine learning, cyber security, and Internet-of-things, which they will use during the last 4-weeks of research phase. In this training phase, we incorporate several group-based hands-on activities that will help teachers to collaborate and share their computer science and education-related knowledge.

To have quality curriculum design, there must first exist a balance between the content and the process, especially where it relates to understanding, or what might also be called critical thinking. In this approach, there are two main goals in a systems design approach to developing a curriculum. The first area is to help students to develop process and abilities and skills increasingly as they progress in grades K-12, or what Jerome Bruner termed the “spiral curriculum” (Bruner, 1991). The second main point is that there is a great need to ensure that students develop conceptual understandings that are rich in what Bruner terms “critical content knowledge” (Bruner, 2006).

In developing a curriculum pragmatically with teachers, we place the emphasis on knowledge and skill development, which is exactly where standards for education put concentration. Based on the scope and sequence of a given district’s plan for student success in each STEM subject, the process moves from concepts to activities by achieving objectives that align with the goals of the enduring knowledge related to cybersecurity and data science. This is also a place where the secondary education standards, which in the case of our research team includes the Texas Essential Knowledge and Skills (TEKS), can be extremely valuable, as in essence, they define the goals, objectives and concepts that can used to align classroom activities that lead to conceptual understandings and directed content knowledge.

The organizing methodologies employed include constructivism, and by extension, problem-based learning and project-based learning. Constructivism is a valid teaching strategy that employs five basic organizational elements that include engagement, exploration, explanation, elaboration and evaluation. Constructivism can be characterized as a five-phased process known as the 5Es, in which each phase begins with the letter E. The 5Es include the engagement phase, the exploration phase, the explanation phase, the elaboration phase and the evaluation phase. As a pedagogical strategy, it allows educators a process by which to facilitate learning opportunities for students. The main premise of this approach is that learners need to take responsibility for their learning and that they learn by being involved in active strategies that require them to problem solve and think critically.

To guide learners to advancement in critical thinking concerning given concepts or topics, the educator facilitates the learning process and the constructivist method with its 5Es becomes an organizational pathway for curriculum development and delivery. The tasks that learners perform need to be organized from the most fundamental to the most complex concepts, and tie directly to real world circumstances. The connections learners make and the knowledge they gain should allow them to address misconceptions they may have, and through their experiences, create new schemas for understanding that bring them to a deeper and broader knowledge that is both practical and functional in their everyday lives.

Case Study 1 – Algebra Concepts and Machine Learning

One research project in which a teacher team investigated had to do with a machine learning program called Correctional Offender Management Profiling for Alternative Sanctions or COMPAS. This software program uses data from criminal offenders to predict the potential of reoffending by analyzing data associated with criminal history, charge degree, race, age, gender and demographics of individuals aggregated as various groups. The machine learning aspect has great consequences, as the outcome can be used to retain or release defendants prior to their trials and for parameters related to sentencing, probation eligibility and treatment programs (Varshney, 2022). The teacher team, both of whom are mathematics teacher at an Early College High School in El Paso, Texas investigated the outcomes of this program and the consequences as it related to various demographic groups, and through their analysis, the began to see that COMPAS was misused and unfairly gave longer sentences to individuals with higher COMPAS scores (Angwin, et. al, 2016).

Based on their analysis of the intersectional fairness of age and race by comparing the age categories of less than 25 and greater than 45, and the races of African American and Caucasians as it related to the charge of battery, discrepancies were found (National Criminal Justice Reference Service, 2022). The first one investigated Caucasians who had committed battery with zero prior counts and their decile scores. The second one investigated African Americans who had committed battery with zero prior counts and their decile scores. The results showed an inherent bias from the machine learning program in the use of the available data, and as a result, there was a regular negative recommendation for younger African American individuals compared to older Caucasians regardless of actual criminal past or crimes.

How this can be integrated into a mathematics classroom at the high school level is also quite creative. For example, in an Algebra I class, there is a section that focuses on linear regressions, which is a fundamental concept for all students to master. This concept is also related to computer science, wherein the relation of machine learning and the analysis of data sets as it relates to the linear regressions through scatterplots. A teacher can relate that data science and machine learning tools utilize algorithms to generate results for linear regressions and produce scatterplots for large data sets that might be too large to do by hand or with the aid of a calculator. The goal is to integrate ideas of analysis to have students make their own predictions and to gauge the viability of decisions generated through the scatterplots, and ultimately, the recommendations made by the data analysis. For many students, where they may have had limited experience with large data sets, can now have a better understanding that the data they provide to the tool may be modified as an output depending on the methodology of the program itself, in other words, that there may be a need to provide options that better fit the expected output. This shows the basics of machine learning and students regular experience with it within the context of a fundamental mathematical concept.

As such, the use of data science as it relates to machine learning will continue to be refined with each iteration and updated data sources that come from the iterative process of intelligent software. The hope is that as time and development progresses, tools such as COMPAS will become more effective and the fairness factor will become closer to being impartial over time due to the integration of large data sets that both represent the correct state of affairs for cause and effect relationships, as well the ways in which the data sets are integrated into the machine learning programs by analysis and evaluation.

Case Study 2 – Interdisciplinary Science and Data Science

A second research project had the teacher-researchers focusing on insider attacks within an organization, wherein a hacker can infiltrate a network. Traditional security measures use passive defense strategies to address cybersecurity issues. The need for more robust measures is being discovered in the field of Psybersecurity, which combines expertise from computer science, cognitive science, psychology and cognitive modeling. The approaches use novel frameworks that utilize a process of models replicating human behavior to help design security measures to address flaws in securing computer systems and networks (Glasser, J., & Lindauer, B., 2013). Cybersecurity measures need to be tailored to the new digital landscape, and Instance-based Learning (IBL) can be used to address these issues

Fostering students' creativity and innovation by presenting opportunities to design, implement, and present meaningful programs through a variety of media is a pivotal way of extending data science concepts into the classroom. Students will collaborate with one another, their instructor, and various electronic communities to solve the problems presented throughout the course. Through computational thinking and data analysis, students will identify task requirements, plan search strategies, and use computer science concepts to access, analyze, and evaluate information needed to solve problems.

In the use of Cognitive modeling in Psybersecurity, it uses large data sets to predict human behavior by simulating through scenarios how people might act in various situations in numerous fields such as education, finances, and healthcare. This process provides insights into how people think and make decisions, as well as emphasizing human and machine learning interactions, wherein the scenarios model the behaviors of cyber attackers, users or defenders (Cranford, et. al. 2020). As a result of this approach, improved decision-making models and algorithms help in the design of more effective defensive counter measures. In terms of the classroom activities, the emphasis is on the importance of making clean coding to make the decisions work, and if not, to correct and review the coding and logic statements. For computer science, the focus is on data science as it relates to working through large data sets of freely available mock demographic information. This skill can then be extrapolated to other data sources, such as sports, entertainment, state or national data repositories. The goal is to utilize data science and logic statements as a basis for analysis to make their own predictions and to gauge the viability of decisions generated through this approach.

By using computer science knowledge and skills that support the work of individuals and groups in solving problems, students will select the technology appropriate for the task, synthesize knowledge, create solutions, and evaluate the results. Students will learn digital citizenship by researching current laws, regulations, and best practices and by practicing integrity and respect. Students also gain an understanding of the principles of computer science through the study of technology operations, systems, and concepts. One method of cognitive modeling involves using a Stackelberg Game and Instance-Based Learning. The purpose of these is to maximize security with limited resources and to use past instances to influence future actions (Maqbool, et. al. , 2022) We can construct computer models to help us view virtual possibilities to influence real-world decision-making. With a large data set (1500 entries for example), rather than having to enter coefficients manually on a calculator or a spreadsheet, the data can be analyzed by coding in order to produce a result based on the parameters that are to be analyzed.

In the classroom, the teacher frames the learning with a question-based approach, through a facilitative style in which they walk around the room to interact better with individual students. The teacher can model the activity for the students from the smart board projector to show the process to the full class and allow the instructions for using Excel to lead the students to produce a decision for each case that is done independently. The teacher should use questions to direct the students so that they can explain the method of calculating within the data set of their choice and to focus the students on the use of calculations to develop formulas that they can also translate to large data sets of various content information.

For students, this leads to developing an understanding of cyber threats within a network, as well as seeing how cognitive modeling & human behavior impact the ability to both combat and address a cyber-attack through data science analysis and synthesis. Finally, students develop skills to write computer science code to enhance instance-based models. The direct skills in data science that students develop include writing basic code and learning the fundamentals of basic coding along with using a referenced library (Lejarraga, T., Dutt, V., & Gonzalez, C., 2012). Students also use artificial intelligence to help generate and debug their coding. As students grow in the ability to manipulate the data, the focus changes to data analysis and creating visual representations of the large data sets, which also provides an intersection for computer science with mathematics.

The experience of taking a model and manipulating parameters was unique to see that computers can perform tasks in a loop to replicate human decision-making. We learned several mathematical equations that demonstrate the concepts we discussed in the introduction. Further research will include implementing the calculation of similarity and replicating the usage of a signal as done in the research done before the project. Overall, we see the importance and variety of factors that go into designing cybersecurity systems for the future to ensure their safety from hackers

Conclusion

By engaging K-12 STEM teachers in an immersive research-based experience with multiple computer science tutorials, problem solving, and hands-on sessions, the teachers work to integrate their new knowledge into a curriculum unit to be integrated into their classrooms in the coming academic year. Curriculum materials that are created by the teachers are posted for dissemination on the program's web site. These curriculum thematic units, in combination with advanced and dedicated mentoring of teachers, define critical elements that will underpin a sustained teaching effort in the teachers' classrooms, and bring data science, cybersecurity and artificial intelligence research to secondary classrooms across our region and beyond.

With data science as an active part of a spectrum of computer science topics that include artificial intelligence and cybersecurity, this professional development program provides a computer science focused research experience for teachers that builds on the proven merits of data science through artificial intelligence and cybersecurity content, provides cooperative linkages between teachers, university faculty, graduate students and external subject matter experts, and integrates educator learning with curriculum development.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias. ProPublica. Retrieved July, 2023, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Bruner, J. S. (1991). The narrative construction of reality. *Critical Inquiry, 18*(1), 1-21.

Bruner, J. S. (2006). *In search of pedagogy volume I: The selected works of Jerome Bruner, 1957-1978*. Routledge.

Cranford, E. A., Gonzalez, C., Aggarwal, P., Cooney, S., Tambe, M., & Lebiere, C. (2020). Toward personalized deceptive signaling for cyber defense using cognitive models. *Topics in Cognitive Science, 12*(3), 992-1011.

Glasser, J., & Lindauer, B. (2013). Bridging the gap: A pragmatic approach to generating insider threat data. *n 2013 IEEE Security and Privacy Workshops* (pp. 98-104). IEEE.

Larson, J., Angwin, J., Kirchner, L., & Mattu, S. (2016). How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Lejarraga, T., Dutt, V., & Gonzalez, C. (2012). Instance-based learning: A general model of repeated binary choice. *Journal of Behavioral Decision Making, 25*(2), 143-153.

Maqbool, Z., Pammi, V. C., & Dutt, V. (2022). Computational modeling of decisions in cyber-security games in the presence or absence of interdependence information. In *Cybersecurity and Cognitive Science* (pp. 357-370). Academic Press.

National Criminal Justice Reference Service. (2022). *Battery*. https://www.ncjrs.gov/

The Importance of Computer Science Education. (2023). *CS1C*. https://sites.uci.edu/cs1c/importance-of-computer-science-education/

Varshney, K. R. (2022). *Trustworthy Machine Learning*. Independently Published.