2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC
Host Sponsor UNC Charlotte
Paper ID #50386

The Kansas Data Science Education Atlas

Srisurya Subhang Chandramouli, Kansas State University
Dr. Safia Malallah, Kansas State University

Dr. Safia Malallah is a teaching assistant professor at Kansas State University, where she completed her Ph.D. in Computer Science and early childhood. Her research is dedicated to advancing computer science and data science education across the PreK-12 and undergraduate levels. Dr. Malallah is particularly passionate about designing innovative and accessible learning experiences that cultivate essential computational skills in students

Lior Shamir, Kansas State University

Associate professor of computer science at Kansas State University.

Bharaneeshwar Balasubramaniyam, Kansas State University

The Kansas Data Science Education Atlas

Srisurya Subhang Chandramouli*, Kansas State University, sschandr@ksu.edu
Bharaneeshwar Balasubramaniyam, Kansas State University, bharanibala@ksu.edu
Safia Malallah, Kansas State University, safia@ksu.edu
Lior Shamir, Kansas State University, lshamir@ksu.edu

Abstract

Recently, Data Science (DS) and Artificial Intelligence (AI) have emerged as key drivers of economic competitiveness, yet access to DS/AI education in rural areas like Kansas is highly uneven. This study introduces the Kansas Data Science Education Atlas, which maps DS/AI offerings across all 105 counties and examines how educational, socioeconomic, and digital factors relate to program availability. Results show strong concentration of programs in counties with four-year universities (Lawrence, Manhattan, Wichita), while 88 counties offer none. Correlation and Machine Learning (ML) analyses confirm that institutional presence, not economic conditions or broadband access, has the strongest correlation with program availability. Random Forest achieved the highest accuracy (91%), with feature importance rankings corroborating the correlation findings and institutional presence consistently dominated economic and digital predictors. Findings suggest that expanding DS/AI education requires targeted institutional investment, faculty development, regional partnerships, and flexible delivery models rather than broadband-focused interventions.

Introduction

The rapid expansion of data-driven technologies has positioned DS and AI as critical workforce competencies [1, 2, 3], with DS employment projected to grow 36% from 2021 to 2031 [4]. Persistent geographic disparities limit equitable access to these fields; rural students remain 23% less likely to access advanced STEM coursework [5, 6]. Although K-12 data literacy efforts are emerging [7], comprehensive geographic analyses of post-secondary DS/AI program distribution are still lacking.

Kansas provides a strong case study: roughly 30% of its students attend rural schools [7], and its institutional diversity and predominantly rural geography highlights how educational resources cluster unevenly. Understanding factors such as broadband access, income, and population density is essential for designing effective interventions, especially as federal initiatives like NSF's Data Science Corps [8, 9] and CDC's workforce programs [10] rely on foundational educational infrastructure that remains uneven. While prior works [11, 12, 13] examine and provide valuable insights regarding education and workforce transformation on a global scale, they fail to address regional nuances. This creates an implementation gap between resource-rich hubs and rural areas [14] such as Kansas. To address this gap, we propose the Kansas Data Science Education Atlas, (1)

*Corresponding Author

to map DS/AI offerings to each Kansas county using correlation techniques and (2) to perform exploratory ML analyses. Based on the hypothesis that geographic and socioeconomic factors serve as primary predictors for educational access, this study aims to answer the four key research questions:

- **RQ1:** How do population density and demographic features correlate with DS/AI educational offerings?
- **RQ2:** How does the geographic distribution of K-12 schools compare to DS program availability, and where are pipeline gaps?
- **RQ3:** How is online DS program availability correlated with geographic inequality when accounting for variations in digital infrastructure?
- **RQ4:** Is there a correlation between county-level economic strength and DS/AI program presence?

Background and Related Work

Preparing a workforce for the Kansas landscape requires updated K-12, community college, and university curricula [15], along with teacher training and extracurricular initiatives such as hackathons and student-led groups like ACM SIGAI [16, 17, 18]. Access to DS education remains uneven, as rural and socioeconomically disadvantaged students have fewer opportunities to engage in advanced STEM courses [19]. These disparities are central for Kansas, motivating this Atlas to map county-level DS/AI gaps and explore correlational patterns between local characteristics and program availability. Causal mechanisms behind these inequities require future longitudinal study.

Geographic disparities also reflect structural injustices. Rawls' Fair Equality of Opportunity stresses equal chances regardless of background [20], while Bourdieu's theory of cultural capital shows how educational systems reproduce advantage [21]. Research consistently finds rural students less likely to pursue postsecondary STEM due to limited coursework and teaching capacity [19], and international studies report similar rural–urban gaps in Norway, England, and Canada [22]. Hillman's concept of "education deserts" shows that many U.S. regions lack nearby colleges or adequate broadband, limiting higher-education access for millions [23]. Spatial inequality is reinforced by uneven institutional distribution [23]. Prior work such as Education 5.0 [11] focuses on emerging fields like AI and Virtual Reality (VR), and the transition from education to workforce has been examined through the lens of Education 5.0 [12, 14]. While the AI implementation gap on a global scale has been addressed [13], there is a critical need for granular county-level analyses to identify disparities such as those addressed by the Kansas Data Science Education Atlas.

Methodology

This study employed a quantitative research design following a modified Element61 DS Methodology framework [24] to systematically document and analyze DS/AI educational offerings across Kansas. As shown in Figure 1, the research was comprised of five phases.

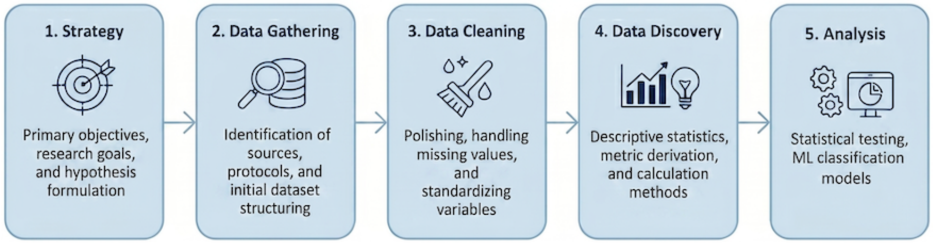

Figure 1: Methodology Workflow.

(1) Strategy

The primary objective was to create a comprehensive geographic atlas documenting DS/AI education availability across Kansas. The aim was to map the spatial distribution of institutions, identify correlations between demographics and educational offerings, analyze digital infrastructure, and examine relationships between county-level economic indicators and program availability.

(2) Data Gathering

The Atlas Dataset serves as the final integrated dataset, consolidating 32 features at the county level for all 105 Kansas counties. It synthesizes variables from Datasets 1–6 (see Table 1). The primary purpose of this dataset is to provide the binary target variable (i.e., *Has Programs*) used for ML classification and performance evaluation. See Appendix B for a full list of abbreviations and acronyms used throughout the paper.

Table 1: Dataset Summary.

Dataset	Category	Source(s)	Key Details	Collection Method
1–2	Educational Institutions	NCES (CCD & IPEDS), IPUMS NHGIS	Includes 1,291 K-12 schools, 76 colleges, and 2019–2023 ACS population data. Dataset 2 aggregated by county-city-zip.	(1) Merged, (2) Derived
3–4	DS/AI Course Inventory	College catalogs via ChatGPT/Claude	106 unique, manually verified courses identified using keywords like 'Machine Learning' and 'AI'.	(3) Scraped, (4) Derived
5	Digital Infrastructure	FCC Broadband Map, IPUMS NHGIS	Covers fixed broadband (25/3 Mbps) and internet adoption. Fixed broadband was prioritized for stability.	Merged, Derived
6	Socioeconomic Indicators	IPUMS NHGIS	County-level data including income, poverty, and unemployment. Features reduced from 30 to 10 for optimization.	Merged, Derived

In the Collection Method column, the terms *Scraped*, *Merged*, and *Derived* indicate respectively that data were obtained through web-scraping or AI-assisted catalog extraction, integrated from multiple official datasets into a unified dataset, or generated as new features constructed from merged fields through cleaning and aggregation.

(3) Data Cleaning

Data consistency was established using SQL within Google BigQuery and Microsoft VS Code. The cleaning protocol involved standardization, categorization, and imputation. County names were normalized to remove nomenclature variations (e.g., unifying "Johnson" and "Johnson County"). For school categorization, a regular expression algorithm mapped institutions to five types (Elementary, Middle, High, Virtual, Other), prioritizing the highest educational level in overlapping cases (e.g., classifying "Jr/Sr High" as High School). Null values were encoded as "DNE" for categorical variables and -1 for numerical variables. To prevent data duplication during joins, aggregation utilized the SUM function for school counts and the MAX function for population and college-type fields.

(4) Data Discovery

Data discovery was guided by the research questions. For RQ1 (demographic correlations), population and income distribution charts were created. RQ2 (pipeline gaps) prompted the mapping of educational institutions and generation of course count rankings. RQ3 (digital infrastructure) resulted in broadband metric summaries, while RQ4 (economic factors) led to correlation matrices pairing socioeconomic indicators with program availability. Heat maps visualized geographic concentration of educational impact, and distribution charts categorized counties by impact levels.

(5) Analysis

Descriptive statistical methods were used including correlation matrices, p-value testing for significance, and Eta-squared analysis for effect sizes. ML classification was included for two purposes beyond the correlation analysis: (1) to test whether the correlational patterns hold under a predictive framework on held-out data, and (2) to identify counties whose feature profiles resemble program-hosting counties despite currently lacking programs, providing a data-driven basis for prioritizing intervention sites.

For ML classification, the binary target (i.e., *Has Programs*) was modeled using 26 predictors. Random Forest, Support Vector Machine (SVM), and Naive Bayes were evaluated through 10-fold cross-validation, measuring accuracy, recall, precision, F1-score, and kappa. Feature importance was determined through Information Gain ranking. ML experiments were conducted in WEKA 3.8 and visualizations were generated in R. The correlation matrices and ML classification use different subsets of the 32 dataset features for distinct reasons (e.g., data leakage); see Appendix A for a full breakdown.

Results

DS/AI Course Offerings and Impact Scores

The impact score is used to quantify the presence and influence of DS/AI education across Kansas. The weights are assigned based on researcher judgment about relative depth of instruction: graduate programs weighted highest as they represent the strongest institutional commitment, undergraduate programs next, with individual courses receiving the lowest weights.

$$\begin{aligned}\text{Impact Score} = {} & (\text{Graduate Programs} \times 2.0) + (\text{Undergraduate Programs} \times 1.0) \\ & + (\text{Graduate Courses} \times 0.2) + (\text{Undergraduate Courses} \times 0.1)\end{aligned} \tag{1}$$

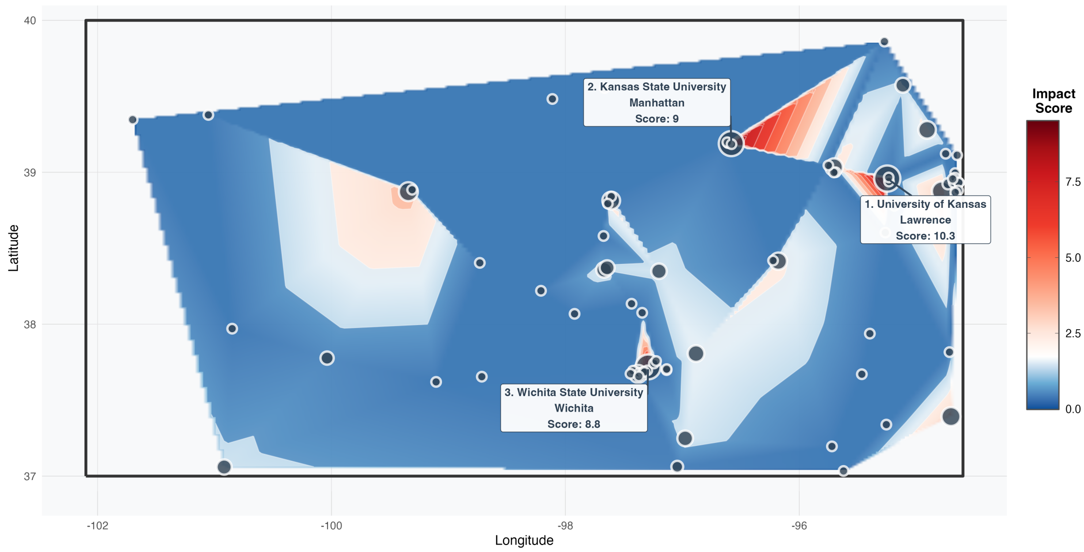

Figure 2: Geographic Intensity of DS/AI Education in Kansas.

Figure 2 shows the interpolated educational impact score (under the researcher-defined weighted scheme described above) across Kansas. Blue and red reflect lower and higher scores respectively to denote the hotspots according to the impact score. The highest-scoring areas correspond to the homes of the three major research universities in Kansas: Lawrence, home of the University of Kansas (10.5), Manhattan, home of Kansas State University (9.8), and Wichita, home of Wichita State University (8.8).

The distribution of program impact reveals stark inequities. Figure 3 shows that 88 counties fall into the "No Programs (0)" category, indicating they have not registered any program or impact. Of the remaining counties, 8 are in the "Low (2–5)" impact range, 6 are in the "Very Low (0–2)" range, and only 3 counties achieved a "High (8+)" impact score, highlighting a significant skew with minimal program presence across most counties.

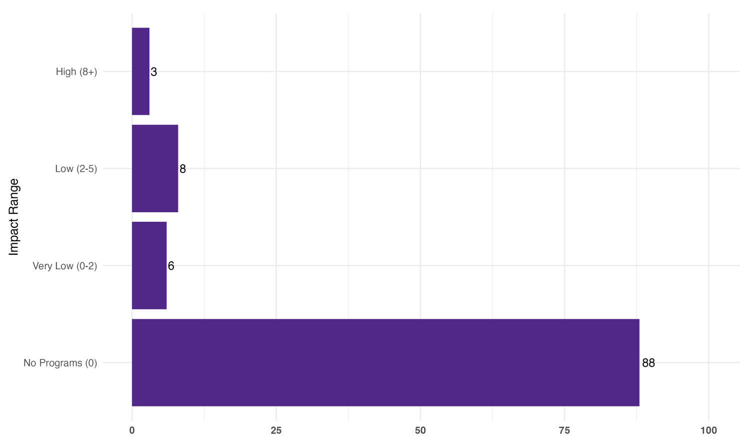

Figure 3: Distribution of Counties by Total Program Impact Score.

Socioeconomic and Demographic Characteristics

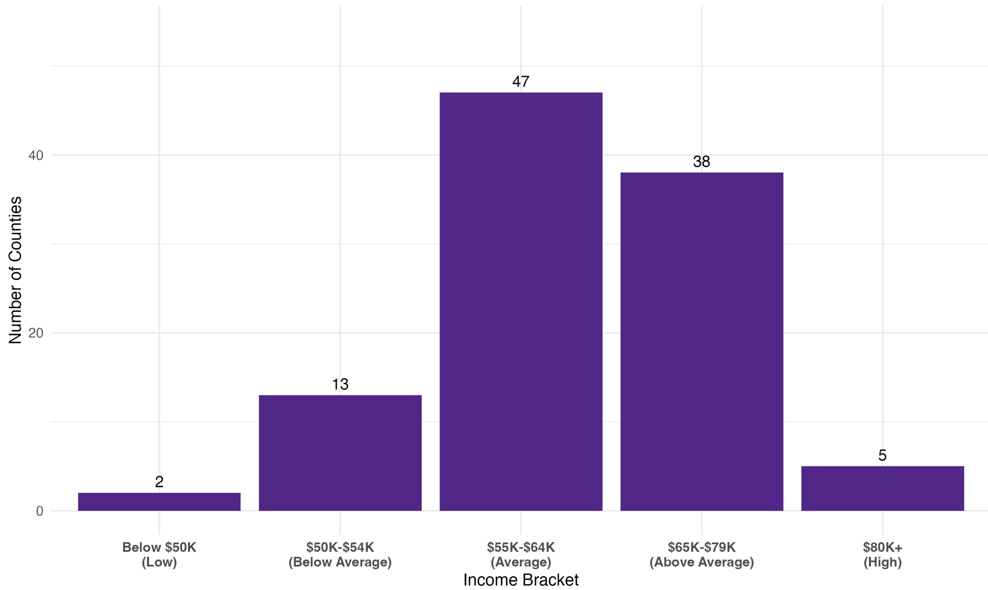

Figure 4: Distribution of Counties by Median Household Income.

Economic indicators show that median household income clusters in average to above-average ranges across Kansas counties. Figure 4 indicates that 47 counties fall into the $55K–$64K (Average) bracket, while 38 counties have median income in the $65K–$79K (Above Average) range. The remaining counties are distributed at the lower end (13 in the $50K–$54K bracket and 2 below $50K) and the highest end (5 in the $80K+ bracket).

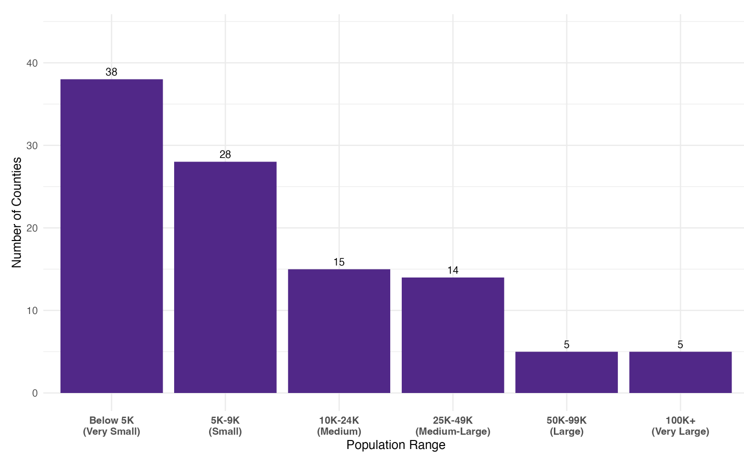

Figure 5: Counties by Population Size.

The population distribution reveals a predominantly rural state. Figure 5 shows that 38 counties fall into the “Below 5K (Very Small)” category, with 28 counties in the “5K–9K (Small)” range. Progressively fewer counties occupy larger population ranges, with 15 in the “10K–24K (Medium)” range, 14 in the “25K–49K (Medium-Large)” range, 5 in the “50K–99K (Large)” range, and only 5 in the “100K+ (Very Large)” range.

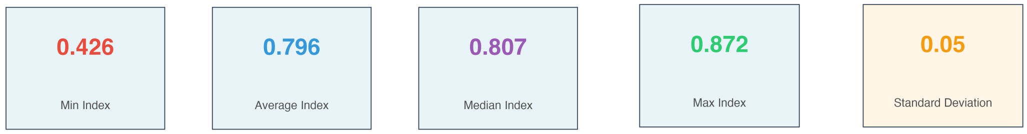

Figure 6: Broadband Access Index Statistical Summary Per County.

Figure 6 provides a concise statistical breakdown of the Broadband Access Index across all counties. The key metrics displayed are the Minimum Index (0.43), Average Index (0.79), Median Index (0.80), Maximum Index (0.87), and Standard Deviation (0.05). These values summarize the central tendency and dispersion of the index scores, indicating that the scores are generally high and tightly clustered, as suggested by the low standard deviation.

ML Classification Performance

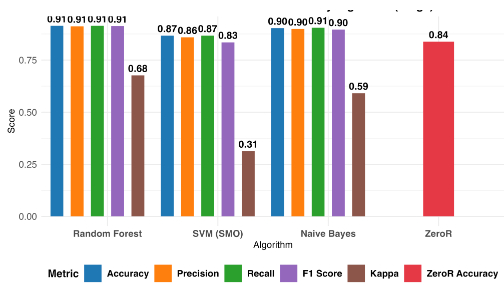

Figure 7: 10-Fold Cross-Validation Metrics.

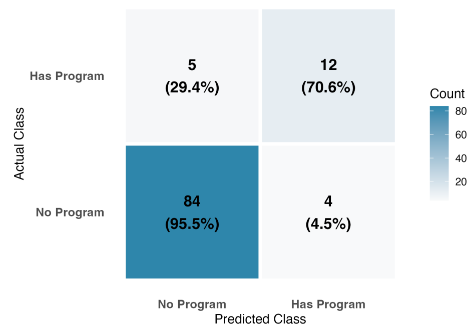

Figure 8: 10-Fold Cross-Validation Random Forest Confusion Matrix.

While ML models showed strong ability in predicting program presence, results must be interpreted cautiously given the class imbalance. As shown in Figure 7, when using 10-fold cross-validation, Random Forest achieved the strongest overall performance compared to SVM and Naive Bayes, with the highest accuracy (0.91), Kappa (0.68), and F1-Score (0.91), effectively identifying minority-class counties. All three models exceeded the ZeroR baseline classification accuracy, which achieved 83.81% accuracy by predicting all counties as having no programs and missing all 17 program-hosting counties. Kappa values (0.31–0.68) confirm that the models learned meaningful patterns beyond simple majority-class prediction, indicating that counties differ not only in DS/AI program availability but also in numerous socioeconomic attributes.

Figure 8 further shows that Random Forest achieved the best balance of predictions (TP = 12, TN = 84), yielding the highest True Positive Rate of 70.6%. It produced more false positives (FP = 4) than SVM (FP = 1) or Naive Bayes (FP = 2) but still provided the strongest overall performance.

Correlation Analysis and Attribute Importance

Correlation analysis confirms the primacy of institutional capacity. As shown in Figures 9 and 10, educational infrastructure variables such as Four-Year Colleges, County Population, and Advanced Degree Rate exhibit the strongest positive correlations with DS/AI program counts ($r = 0.85$–0.95). In contrast, economic indicators such as Median Household Income and Poverty Rate display weak correlations ($r = 0.16$–0.21).

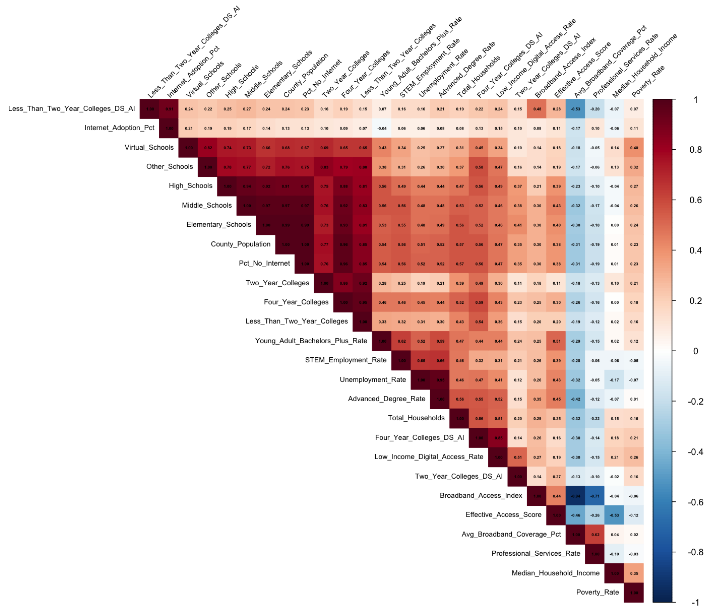

Figure 9: Predictor + DS/AI College Count Correlation Matrix.

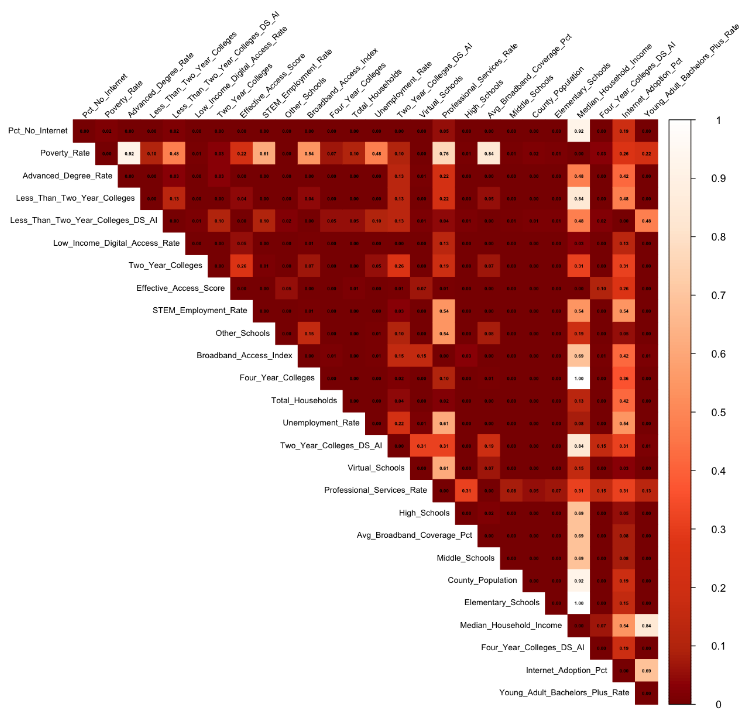

Figure 10: Predictor + DS/AI College Count P-Value Matrix.

Figure 10 also shows that these educational variables are statistically significant ($p < 0.05$) with DS/AI program presence, while most economic variables are not. Digital access metrics, including

Broadband Access Index and Internet Adoption Percentage, likewise show non-significant p-values, reinforcing that broadband uniformity across counties does not explain program distribution.

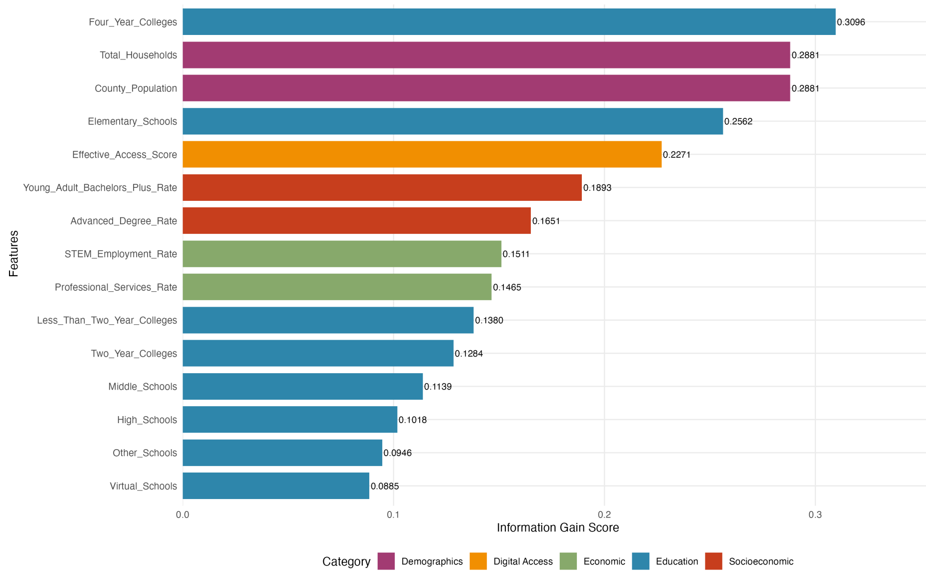

Figure 11: Feature Analysis Ranking.

Figure 11 shows the assessment of predictor importance using Information Gain-based ranking. Four-Year Colleges emerge as the most influential feature, followed by County Population and Advanced Degree Rate, while broadband and income-related variables rank at or near zero. Together, these figures confirm that institutional capacity, not economic wealth or digital infrastructure, is associated with DS/AI program availability in Kansas.

Educational Infrastructure and DS/AI Course Concentration

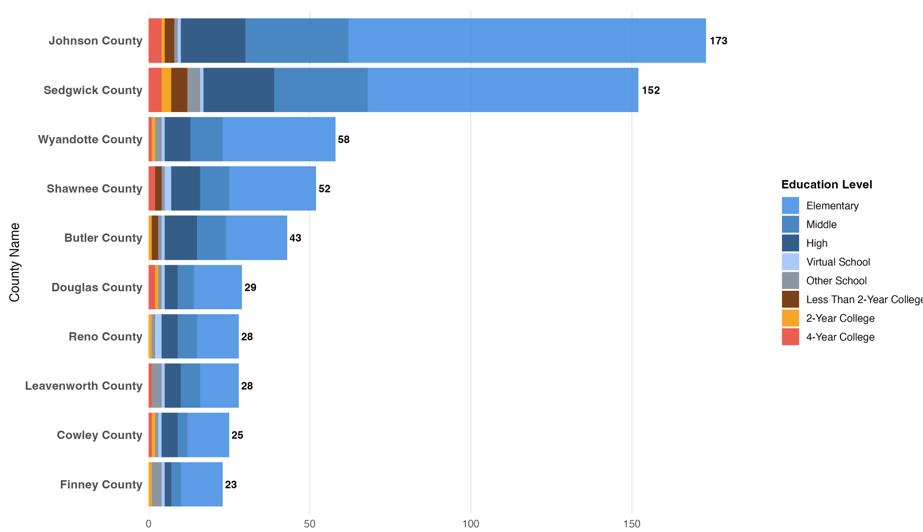

Figure 12: Top 10 Counties by Total Number of Educational Institutions Denoted by Education Level.

The analysis of educational institutions across Kansas reveals significant geographic concentration. Figure 12 displays the distribution of educational institutions across the 10 Kansas counties with the highest total count, segmented into eight distinct educational levels: Elementary, Middle, High, Virtual, Other, 2-Year College, 4-Year College, and Less Than 2-Year College. The visualization highlights counties with a high density of educational resources.

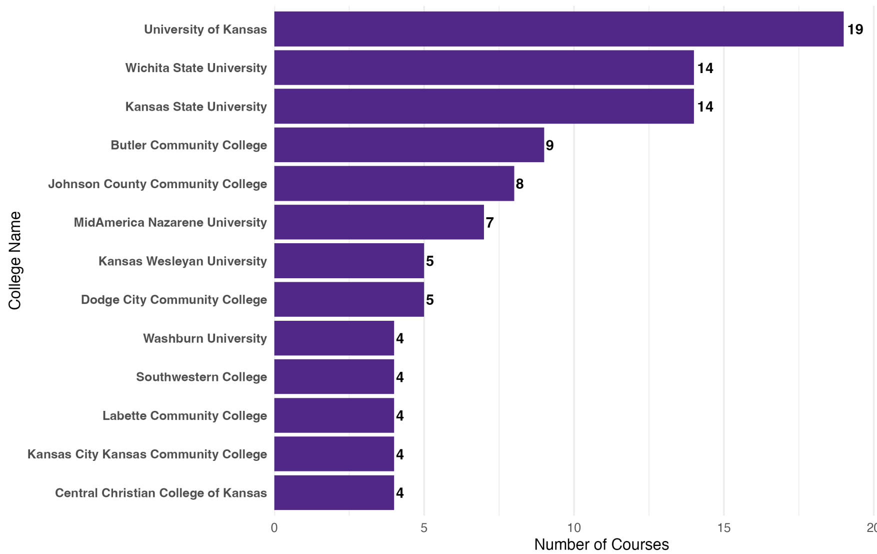

Figure 13: Top 13 Colleges by Course Count.

Examining DS/AI course offerings specifically, Figure 13 shows that the University of Kansas leads with 19 courses, while Kansas State University and Wichita State University each offer 14 courses. The remaining institutions show smaller course counts, ranging from nine courses at Butler Community College down to four courses for the bottom four ranked schools, indicating a clear hierarchy in DS/AI course offerings.

Discussion

RQ1 – Demographic Features and Educational Equity.

The strong correlation between population density and DS/AI program availability ($r > 0.90$, $p < 0.05$) (Figures 9–10) reveals more than a simple urban-rural divide. Figure 2 shows programs concentrated in three university towns (Lawrence, Manhattan, Wichita), while Figure 3 indicates 88 counties have no program impact. This clustering is consistent with a systemic pipeline gap: counties with comprehensive K-12 infrastructure (Figure 12) also tend to host research universities, with complete educational pathways present only those geographies. The 38 counties with populations below 5,000 (Figure 5) show no DS/AI program availability in this dataset. Prior literature suggests that rural students may also have fewer earlier exposure points (e.g., AP Computer Science, robotics clubs) [16, 17, 18], though this study does not measure K-12 course offerings directly.

The ML analysis identifies four False Positive counties (Figure 8), counties whose demographic and structural profiles match program-hosting counties, yet currently offer no DS/AI programs. These cases are notable because the model's misclassification itself carries meaning: it suggests these counties possess the preconditions associated with program availability in other Kansas counties. This makes them potentially higher-leverage targets for intervention than counties whose profiles differ markedly from program-hosting ones. Policymakers may conduct needs assessments in these specific counties to identify barriers (faculty recruitment, startup funding, accreditation support) and target resources accordingly.

RQ2 – Educational Infrastructure and Pipeline Gaps.
Figure 11 shows that Four-Year Colleges dominate with the strongest feature importance ranking (Information Gain = 0.352). The high correlation between Four-Year Colleges and both Elementary Schools ($r = 0.96$) and Advanced Degree Rate ($r = 0.95$) shown in Figure 9 reveals that successful DS/AI programs are associated with multi-level capacity: counties where K-12 schools and universities co-exist show higher program availability. This pattern is consistent with an ecosystem dependency that may help account for why the 88 counties without programs (Figure 3) face more than a simple program gap. They lack interconnected educational infrastructure that enables program sustainability. The co-location pattern in Figure 13 (University of Kansas: 19 courses; Kansas State and Wichita State: 14 each) demonstrates programs cluster where comprehensive educational systems already exist.

From the findings, potential strategic approaches can include a regional hub model to serve False Positive counties identified in Figure 8, offering synchronous online instruction during semesters and in-person intensive sessions during summers, or expanding 2+2 articulation agreements to leverage two-year colleges by creating guaranteed transfer pathways to research universities for DS/AI degree completion. The data does not directly test the efficacy of any intervention and warrants further study. The causal direction remains unclear from these cross-sectional data: do universities create local STEM capacity, or do they locate in counties with pre-existing capacity? This question shapes whether intervention should prioritize faculty recruitment (supply-side) or student demand cultivation (demand-side), requiring longitudinal analysis to resolve.

RQ3 – Digital Infrastructure and Geographic Inequality.
Broadband infrastructure shows minimal predictive power despite widespread coverage (Figure 6). The near-zero Information Gain for digital metrics (Figure 11) and weak correlations ($r < 0.12$, $p > 0.05$) (Figures 9–10) reveal that adequate infrastructure exists, yet programs remain concentrated in university towns. Broadband access is not significantly correlated with local program availability in this dataset, suggesting it may be more relevant to program consumption than program creation, though this distinction requires further study. Three factors explain this limitation. First, counties without universities cannot recruit PhD-level faculty remotely. Second, courses involving robotics, IoT sensors, or hardware-software integration require physical laboratory infrastructure that broadband alone cannot provide. Third, establishing new programs requires institutional processes (e.g., curricular approval, accreditation, budgeting) independent of technical connectivity.

These results suggest that leveraging existing infrastructure through hybrid delivery models is a viable strategy. Synchronous cohort models, where students across multiple counties take coordinated online courses with periodic meetups, may maintain peer learning while distributing access more equitably. The cross-sectional design cannot determine whether historical broadband deficits correlate with lower program counts, or whether broadband access was unrelated to program distribution. Additionally, the 25 Mbps/3 Mbps threshold may prove insufficient for emerging applications (large dataset transfers, cloud GPU computing), suggesting that current "adequate" infrastructure may require upgrading within 5–10 years.

RQ4 – Economic Features and Program Presence.
Economic indicators show weak correlation (e.g., Median Household Income: $r = 0.16$–0.21, $p > 0.05$) (Figures 9–10), and near-zero feature importance (Figure 11). Despite 85 counties in average-to-above-average income brackets (Figure 4), most have zero program impact (Figure 3). The weak correlation between economic indicators and program availability suggests that income alone may not be the primary barrier to program expansion, which is a finding consistent with supply-side explanations (e.g., institutional presence, teaching capacity). Interventions such as faculty fellowship programs offering stipends to DS/AI specialists committed to underserved institutions, startup grants for community colleges in False Positive counties covering initial hiring and infrastructure, and stackable credential pathways, where certificates ladder into degrees, reducing financial barriers while ensuring transfer credit guarantees may help bring a positive impact. While economic indicators weakly predict program availability, they may moderate program success once established. Counties with higher median incomes may have students better able to afford tuition and persist through challenging coursework. This analysis examines availability and not outcomes, a limitation requiring longitudinal student success data to address.

Conclusion, Limitations, and Future Work

This study maps DS/AI education across Kansas and finds that programs are concentrated in a few four-year universities, leaving most counties without access. Institutional capacity, especially the presence of four-year colleges, and population density are highly correlated with program availability, while broadband and economic factors show little correlation. ML models reached up to 91% accuracy, reinforcing these patterns. This has implications for how expansion efforts might be prioritized in ways like targeted institutional investment, faculty development, and regional partnerships rather than infrastructure-focused interventions. The four false-positive counties represent promising sites for pilot expansion efforts. Though specific intervention designs require further longitudinal and experimental study. The public dashboard showcasing the data and insights is accessible at https://atlas-frontend-44fq.onrender.com/.

This study has several limitations. The current manual course collection represents a single snapshot in time. To rectify this, future work might involve developing automated scraping protocols to provide real-time updates via a public dashboard. Moreover, we acknowledge that the Impact Score weights require empirical validation. Alternative weighting schemes (e.g., equal weights, faculty survey data) could yield different county rankings. The binary target variable used in ML classification (Has Programs) is less sensitive to this limitation than the continuous score, as it depends only on whether any program or course exists, not on the specific weights. We intend on conducting deeper scholarly research and sensitivity testing on the weights. To overcome the statistical constraints of class imbalance and the small sample size of a single state (i.e., Kansas), expanding to a multi-state regional analysis could be a possible area of exploration. Finally, the cross-sectional design of this study prevents causal conclusions. Therefore, this study would become a five-year longitudinal study to observe whether institutional investments or demographic shifts serve as the primary catalyst for new DS/AI programs.

Acknowledgment

This work was funded by the National Science Foundation (NSF) with Grant No. 2148878. However, the authors alone are responsible for the opinions expressed in this work and do not reflect the views of the NSF.

References

[1] Holistic Data Science. Global Economic Impact of Artificial Intelligence: Horizon 2040. https://holisticds.com/en/articulo/ai-economy/. Accessed: January 10, 2026.

[2] World Economic Forum. Data Science in the New Economy: A New Race for Talent in the Fourth Industrial Revolution. In *World Economic Forum Annual Meeting*, volume 1, pages 1–22, 2019.

[3] Harvard John A. Paulson School of Engineering and Applied Sciences. What is Data Science? Definition, Skills, Applications & More. https://seas.harvard.edu/news/what-data-science-definition-skills-applications-more. Accessed: January 10, 2026.

[4] R. Farrell and M. Hussain. Data occupations with rapid employment growth, projected 2021–31. *Career Outlook*, 2023.

[5] Melissa Emrey-Arras. STEM Education: Selected Federal Initiatives, Challenges, and Approaches to Supporting Rural Populations. Report to Congressional Committees. GAO-25-107371. *US Government Accountability Office*, 2025.

[6] Yi Thung Gan, Kher Hui Ng, Tissa Chandesa, Xun Yu Chai, Alberto Luis August, and Kean Wah Lee. A systematic review of STEM interventions in rural education: July 2013 to june 2023. *Journal for STEM Education Research*, pages 1–41, 2024.

[7] Safia A. Malallah, Lior Shamir, William Henry Hsu, Joshua Levi Weese, and Salah Alfailakawi. Data Science (Dataying) for Early Childhood. In *2023 ASEE Annual Conference & Exposition*, 2023.

[8] National Science Foundation. NSF 24-560: Data Science Corps (DSC). https://new.nsf.gov/funding/opportunities/data-science-corps-dsc, 2024. Accessed: January 10, 2026.

[9] National Science Foundation. Data Science Corps Awards Aim to Develop Workforce-Ready Data Scientists. https://www.nsf.gov/news/news_summ.jsp?cntn_id=303629. Accessed: January 10, 2026.

[10] Mary Catherine P. Bertulfo, Robert D. Kirkcaldy, Laura H. Franzke, Sridhar R. Papagari Sangareddy, and Faisal Reza. Advancing data science among the federal public health workforce: The data science upskilling program, centers for disease control and prevention. *Journal of Public Health Management and Practice*, 30(2):E41–E46, 2024.

[11] Shabir Ahmad, Sabina Umirzakova, Ghulam Mujtaba, Muhammad Sadiq Amin, and Taegkeun Whangbo. Education 5.0: Requirements, Enabling Technologies, and Future Directions. *arXiv preprint arXiv:2307.15846*, 2023.

[12] Sharareh Shahidi Hamedani, Sarfraz Aslam, Bha Aldan Mundher Oraibi, Yap Bee Wah, and Shervin Shahidi Hamedani. Transitioning towards Tomorrow's Workforce: Education 5.0 in the Landscape of Society 5.0: A Systematic Literature Review. *Education Sciences*, 14(10): 1041, 2024.

[13] Sourojeet Chakraborty and Daniela Galatro. Artificial Intelligence (AI) Equality in Engineering Education: Strategies to Unite the AI Gap between the Global North & South. In *2025 IEEE Global Engineering Education Conference (EDUCON)*, pages 1–10. IEEE, 2025.

[14] Bianca Ifeoma Chigbu and Sicelo Leonard Makapela. AI in Education, Sustainability, and the Future of Work: An Integrative Review of Industry 5.0, Education 5.0, and Work 5.0. *Journal of Open Innovation: Technology, Market, and Complexity*, page 100645, 2025.

[15] Institute of Education Sciences, U.S. Department of Education. Data Science Education. https://ies.ed.gov/use-work/supporting-recovery-with-evidence-based-practices/data-science-education. Accessed: January 10, 2026.

[16] Carrie Grace Aponte, Safia Malallah, and Lior Shamir. A Review of K-12 Data Science Education in the United States: Trends, Tools, and Gaps. In *2025 ASEE Annual Conference & Exposition*, 2025.

[17] The Academic Insights. Analytics and Beyond: Extracurricular Activities That Make Learning Fun. https://theacademicinsights.com/analytics-and-beyond-extracurricular-activities-that-make-learning-fun/. Accessed: January 10, 2026.

[18] ACM SIGAI. About Us. https://sigai.acm.org/main/aboutus/. Accessed: January 10, 2026.

[19] Guan K. Saw and Charlotte A. Agger. STEM Pathways of Rural and Small-Town Students: Opportunities to Learn, Aspirations, Preparation, and College Enrollment. *Educational Researcher*, 50(9):595–606, 2021.

[20] Jon Mandle. *Rawls's 'A Theory of Justice': An Introduction*. Cambridge University Press, 2009.

[21] Pierre Bourdieu. The forms of capital. In John G. Richardson, editor, *Handbook of Theory and Research for the Sociology of Education*, pages 241–258. Greenwood, New York, 1986.

[22] Alexander Zahl-Thanem and Johan Fredrik Rye. Spatial Inequality in Higher Education: A Growing Urban–Rural Educational Gap? *European Sociological Review*, 40(6):1067–1081, 2024.

[23] Nicholas W. Hillman. Geography of College Opportunity: The Case of Education Deserts. *American Educational Research Journal*, 53(4):987–1021, 2016.

[24] Element61. Data Science Methodology. https://www.element61.be/en/competence/data-science-methodology. Accessed: February 17, 2026.

Appendix

A: Feature Set Summary

Feature	**Correlation Matrix (Figures 9–10)**	**ML Classification**
County Name	Excluded	Excluded
County Latitude	Excluded	Included
County Longitude	Excluded	Included
Has Programs	Excluded	Included (target)
Total Program Impact Score	Excluded	Excluded
Online Impact Score	Excluded	Excluded
Four-Year Colleges DS/AI	Included	Excluded
Two-Year Colleges DS/AI	Included	Excluded
Less Than Two-Year Colleges DS/AI	Included	Excluded
All remaining 23 features	Included	Included

Correlation matrices include 26 features (32 total minus 6 excluded). ML classification uses 26 predictors (32 total minus 6 excluded). The two sets exclude different features for different reasons: latitude/longitude and *Has Programs* are removed from visualizations for interpretive reasons, while program-derived counts are removed from ML training specifically to prevent data leakage, as they are direct derivatives of the target variable.

B: Table of Abbreviations and Acronyms

Abbreviation / Acronym	Full Term
ACM SIGAI	Association for Computing Machinery Special Interest Group on Artificial Intelligence
ACS	American Community Survey
AI	Artificial Intelligence
AP	Advanced Placement
CCD	Common Core of Data
CDC	Centers for Disease Control and Prevention
DNE	Does Not Exist
DS	Data Science
DS/AI	Data Science and Artificial Intelligence
FCC	Federal Communications Commission
FN	False Negative
FP	False Positive
GPU	Graphics Processing Unit
IPEDS	Integrated Postsecondary Education Data Systems
IPUMS NHGIS	Integrated Public Use Microdata Series National Historical Geographic Information System
IoT	Internet of Things
K-12	Kindergarten through 12th grade
Mbps	Megabits per second
ML	Machine Learning
NCES	National Center for Education Statistics
NSF	National Science Foundation
PhD	Doctor of Philosophy
RQ	Research Question
SQL	Structured Query Language
STEM	Science, Technology, Engineering, and Mathematics
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
VS Code	Visual Studio Code
WEKA	Waikato Environment for Knowledge Analysis