iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** X **Month of publication:** October 2024

DOI: https://doi.org/10.22214/ijraset.2024.64518

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue X Oct 2024- Available at www.ijraset.com

Data Quality and Data Governance: Investigating the Impact on Data Science Outcomes

DR. Diwakar Ramanuj Tripathi[1], Mansi Mukesh Jain[2], Poonam Gajanan Kharche[3]
[1]*HOD, PG Department Of Computer Science,* [2, 3]*Research Scholars, S.S. Maniar College of Computer & Management, Nagpur*

Abstract: ***The purpose of this study is to evaluate the effects that high data quality and effective data governance have on the outcomes of data science, with a particular emphasis on the vital roles that these factors play in improving decision-making within businesses. The rising use of data science projects by enterprises for the purpose of digital transformation presents major hurdles, particularly with regard to the trustworthiness of data and compliance with appropriate legal frameworks. A high data quality improves model accuracy, decision-making efficiency, and error reduction, while good governance boosts compliance with rules, data security, and accessibility, according to the findings of the study, which were derived from an analysis of data collected from professionals working in data science and related sectors. These findings demonstrate the synergy that exists between data quality and governance, underscoring the fact that businesses may obtain more dependable and trustworthy outcomes by prioritizing these components, which in turn supports digital transformation programs that are effective and responsible.***

Keywords: ***Data Quality, Data Governance, Data Science Outcomes, Decision-Making, Digital Transformation, Data Trustworthiness, Data Security, Accessibility.***

I. INTRODUCTION

Information science drives have filled in prevalence as a device for organizations to support their computerized change in the beyond quite a while. Associations keep on confronting difficulties while attempting to depend on information science results for navigation, regardless of the way that the information is much of the time ineffectively obtained and its consistence with significant lawful structures, cultural standards, and values isn't evident all the time. These vulnerabilities upset the reception and organization of information science arrangements because of the potential for monetary gamble and reputational misfortune. For example, in order for asset managers to feel confident using data science, they need to trust the results while making decisions about the physical assets they are managing. Such matters include, but are not limited to, the timing and location of bridge repairs and highway maintenance. Erring on the side of caution can lead to excessive costs, while reckless maintenance delays can threaten public safety. Thus, for data science to be effectively implemented, businesses must have faith in the reliability of the results. A developing number of organizations are going to information administration for of expanding client certainty. Since it is muddled how information administration adds to the turn of events and support of confidence in information science for navigation, there have been calls for more concentrate around here.

It is the goal of data science to help people make better judgments. Data science is defined by Dhar (2013) as the practice of providing explanations and projections based on information gathered via systematic study. This leads to numerous points of departure between data science and more conventional branches of science. For a long time, scientists have investigated the same subject by gathering data. The next step is to evaluate the data in order to have a deeper understanding of the subject. In contrast, data scientists often compile a variety of pre-existing datasets in search of correlations that disclose surprising or previously unknown valuable insights. Research has shown, however, that risks related to incorrect data interpretation could emerge when analytical approaches take precedence over domain knowledge. One may easily conclude that data scientists' judgments are rational solely due to the automation of the process. However, the quality of the data at hand dictates the decision-making process, as is the case with any other decision-making process, due to constrained rationality. The results produced by data science models are dictated by the data and time that are currently available to them. According to Gama (2013), one example of restricted rationality in data science is the trade-off between the space and time required to answer a query and the precision of the conclusion. A major number of associations are executing information administration to assume command over these variables, which isn't is business as usual. In spite of the fact that information science is broadly perceived as a useful asset for navigation, its viability is restricted by the nature of the models and information utilized.

Information administration is characterized as "the activity of power and control (arranging, checking, and authorization) over the administration of information resources" and it offers immediate and aberrant advantages.

Provide examples of how data governance can change how businesses create, collect, and use data. For instance, data governance has the potential to greatly increase the visibility of data science results for infrastructure management in a smart city context. Yet, data governance efforts driven by IT haven't always been a success story. This is usually because of technical implementation challenges that have to be handled system-by-system. In order to have faith in the offered outcomes, this study offers a fresh approach by focusing on data governance as a data science prerequisite that must be addressed. This paper asserts that in order to have faith in data science outcomes, certain socio-technical constraints, known as boundary requirements, must be satisfied. These requirements, which relate to the "who," "where," and "when" aspects of data science, must be met before the results can be used. Based on previous research, data governance might be considered an essential component of data science. We are fundamentally inspired by what information administration means for the consequences of information science direction. To handle this, two resource the executives related information science contextual investigations were dissected, with an emphasis on how information administration advances trust in information science dynamic results — an essential condition for exact prescient direction. An information science work to work on the proficiency of street fix by prescient upkeep is the main situation being thought of.

II. LITERATURE REVIEW

Micheli, M., et.al., (2020) concentrated on four developing information administration models in the stage society of today. While the prevailing model of corporate stages gathering and productively taking advantage of tremendous measures of individual information is as of now getting a ton of consideration, different members in information administration incorporate little undertakings, government organizations, and municipal society. The paper clarifies four models—public data trusts, data cooperatives, data sharing pools, and personal data sovereignty—that have emerged from the actions of these players. We suggest a framework of data governance that is grounded on social science. We define the models as a result of the roles and interactions between the stakeholders, their articulations of value, and governance principles, drawing on the idea of data infrastructure. We examined the actors' competitive fights for data governance in order to address the politics of data. This paradigm highlights the complex relationships between corporations and other social and economic actors as well as the power dynamics inside the data governance models that are evolving in this mostly corporate-dominated environment. These examples demonstrate the importance of public entities and civic society in democratizing data governance and spreading the value generated by data. The purpose of this paper is to provide guidance for future research on socio-technical imaginaries for data governance, especially in light of the current, highly active policy debate in Europe, by discussing the models, their underlying principles, and their limits.

Janssen, M., et.al., (2020) researched approaches and difficulties related with data administration for such systems, and proposed a design for reliable BDAS data administration. The multiplication of Big, Open, and Linked Data (BOLD) has made ready for Big Data Algorithmic Systems (BDAS), which routinely utilize different types of simulated intelligence, for example, AI and brain organizations. Due to the developing number of situations when these systems are supposed to pursue choices that influence people, networks, and society at large, they are dependent upon severe moral and legitimate prerequisites, and their inadequacies should not go on without serious consequences. However, in any case, they are reliant upon data that isn't simply huge, effectively accessible, and organized, yet in addition shifted, consistently changing, and communicated quickly and progressively. This sort of data is trying to make due. To address these difficulties and take advantage of BDAS chances, associations are progressively getting progressed data administration capacities. The framework level controls, shared proprietorship, self-sovereign personalities, data, cycle, and calculation stewardship, controlled opening of data and calculation to outer examination, and believed data sharing inside and between associations are undeniably advanced by the system. In light of thirteen plan standards, the structure is logically accommodated both a solitary organization and numerous organized endeavors.

Eyal, P., et.al., (2021) utilized a couple of chosen stages (Amazon Mechanical Turk, CloudResearch, and Productive) and boards (Qualtrics and Dynata) to assess basic parts of online social examination data quality. We studied the conduct research local area to distinguish the most basic components for scientists to be aware to decide the main parts of data quality. We found that honesty, reliability, attention, and understanding are the most crucial aspects. In two separate trials with a total of approximately 4000 participants, we examined changes in these data quality attributes with and without the use of data quality filters, namely approval ratings. In instance, there were noticeable differences in honesty, attentiveness, and comprehension across the sites. In the wake of taking a gander at every one of the rules in Study 1 (without channels), they tracked down that main Productive gave data fantastic quality. In Study 2 (with channels), we tracked down that Cloud Research and Productive had exceptionally top notch data.

A horrifyingly low data quality was as yet shown by MTurk even subsequent to utilizing data quality channels. On MTurk specifically, we found that use reason and recurrence, not standing (endorsement rating), anticipated data quality.

The individuals who said they made a large portion of their cash on the site yet put in a couple of hours there each week had the most obviously terrible data quality. We lay forward an arrangement for future examinations that will explore the consistently changing condition of online exploration data quality and look at the different stages and boards in view of these critical measurements.

Yallop, A. C., et.al., (2023) studied an ethical perspective is included in the proposed framework, which goes beyond only following privacy and protection rules. aimed to establish guidelines for ethical data handling in the hotel and tourism sectors, with the hope of fostering effective data governance practices. Big data and analytics are becoming more significant tools for organizations in the tourist and hospitality industries (THOs) to help them make strategic decisions. In times of crisis and uncertainty, data analytics may assist THOs get the insights they need to keep their businesses running and rebuild the tourism and hospitality sectors. While digital technologies and big data are widely recognized as powerful engines of economic growth, they also give birth to concerns regarding privacy, security, and ethics. This article takes an organizational and stakeholder stance through a literature scoping study in order to offer a review of a neglected subject and to direct future research on data ethics and data governance. Additionally, it considers other crucial elements of ethics and privacy, the equitable transfer of traveler data, and the ability of THOs to demonstrate their social license to operate through the development of trustworthy relationships with stakeholders. Since this is one of the first studies to look at how THOs may build an ethical data framework, it lays the groundwork for more theoretical and practical studies on data governance frameworks like these in the future. Not only does it benefit practitioners, but it also adds to the body of knowledge on data governance and ethics in sectors like hospitality and tourism. The reason behind this is that it can serve as a blueprint for data governance processes within enterprises.

Davidson, E., et.al., (2023) investigated the unique issue on data administration, computerized advancement, and grand difficulties features examinations concerning the significance of data administration with regards to endeavors to handle extraordinary issues through the creative utilization of computerized innovation. Data in the present data-rich climate has various purposes, both for development and society all in all, and it likewise has its disadvantages. In any case, it is not ensured that main beneficial things will unfold. Serious worries in regards to fairness, protection, data security, and the probability of data misuse are elevated by the creation and utilization of huge scope data stores. Executing viable data administration techniques inside and across associations is basic for encouraging advanced development while adjusting the cultural, monetary, and innovative benefits and risks for people, organizations, and society at large. To improve computerized advancements for change and social advantage, this presentation paper considers the present status of information systems (IS) research and recommends conceivable future bearings for data administration grant at different levels.

Ravikumar, R., et.al., (2023) delved into the utilization of big data in healthcare for the purpose of knowledge management by means of case studies that employ various analytics, and it draws parallels that can be utilized in the realm of higher education. Everyone agrees that a company can't thrive and grow without knowledge management. Institutions of higher learning, often called "knowledge centers," generate massive volumes of data daily. This data analysis, when coupled with the appropriate computational methods, has the potential to improve both the efficiency of organizations and the quality of education that students get. Healthcare firms create massive volumes of data due to the prevalence of digital technologies used to handle patient records and internal processes. This data has the potential to improve patient health, organizational operations, and the prevention of infectious disease transmission and other negative public health circumstances when used properly. This is where big data analytics come in handy; they provide rational methods for sifting through mountains of data, which in turn aids analysts and companies in making better, faster judgments. Higher education, like the healthcare industry, generates massive amounts of heterogeneous data that mask crucial information. As a result, the education sector can also benefit from the big data-driven performance improvement strategies used by healthcare businesses. As a result, it highlights the possibilities for adapting analytics approaches and tools from the healthcare sector to the academic sector.

III. RESEARCH METHODOLOGY

A. Research Design

Within the scope of this investigation, a quantitative research design is utilized to evaluate the influence that good data quality and efficient data governance have on the outcomes of data science. For the purpose of collecting measurable data, a structured approach was utilized, which made it possible to do statistical analysis on the correlations between the variables.

The purpose of this study is to identify particular areas of data science that are favorably influenced by data quality and governance. The findings of this study will provide empirical proof for organizations who are looking to enhance their data practices.

B. Data Collection

The researchers surveyed professionals in data science, data analytics, and related domains via online survey to compile their findings. Many different methods were used to spread the survey link in order to reach a big number of people. These included relevant industry forums, social media, and professional networks. The respondents were asked to provide direct input on their experiences with data quality and governance in their own businesses, which ultimately resulted in a dataset that was both diverse and comprehensive.

C. Data Analysis

The purpose of the data analysis was to determine the magnitude of the influence that good data quality and efficient data governance have on the outcomes of various components of data science. In this study, descriptive statistics were utilized to quantify the respondents' assessments of improvements across key dimensions. These characteristics were model correctness, decision-making efficiency, data processing speed, error reduction, and compliance with data regulations. The results demonstrated considerable enhancements for high data quality, with 78% of respondents recognizing enhanced model correctness and 70% recognizing higher data usability for predictive analytics. These findings show that the data quality has significantly improved. Furthermore, 68 percent of respondents observed an increase in the efficiency of decision-making, while 65 percent indicated a faster processing of data. 83% of respondents reported greater compliance with data regulations, and 80% of respondents highlighted enhanced data security. Effective data governance had a comparable influence on the percentage of respondents. In the combined examination of the two elements, even more dramatic benefits were discovered: the accuracy of the model increased by 85 percent, the efficiency of decision-making increased by 80 percent, and compliance with rules increased by 88 percent. In order to provide a clear picture of the link between high data quality, effective governance, and superior data science outcomes, these results were visually represented through tables and figures. To emphasize the crucial need of maintaining high standards in data management methods, this correlation was provided.

D. Sample Size

In order to assure statistical significance and the reliability of the results, the study targeted a sample size of 300 respondents. The intended degree of confidence in the results and earlier studies in the field were taken into consideration while determining this sample size. The final dataset included professional replies from a wide range of industries, such as technology, healthcare, and finance, giving a thorough overview of how data governance and quality affect data science outcomes.

IV. DATA ANALYSIS

Strong data quality greatly improves data science results, as seen by the benefits noted in a number of areas. Model accuracy is most significantly impacted; 78% of respondents reported improved predictive model performance. Likewise, 70% of respondents reported better data usability for predictive analytics, highlighting how crucial trustworthy data is to producing insightful analyses. 68% of respondents reported higher decision-making efficiency as a result of less manual data cleaning, and 65% reported faster data processing as a result of fewer data discrepancies. Furthermore, 62% of respondents observed that models reduced errors, suggesting that high-quality data decreases errors. All things considered, these results demonstrate how important data quality is to the optimization of data science processes.

Table 1. High Data Quality's Effect on Data Science Results

Aspect of Data Science	Percentage Reporting Improvement (%)
Model Accuracy	78%
Decision-Making Efficiency	68%
Data Processing Speed	65%
Error Reduction in Models	62%
Data Usability for Predictive Analytics	70%

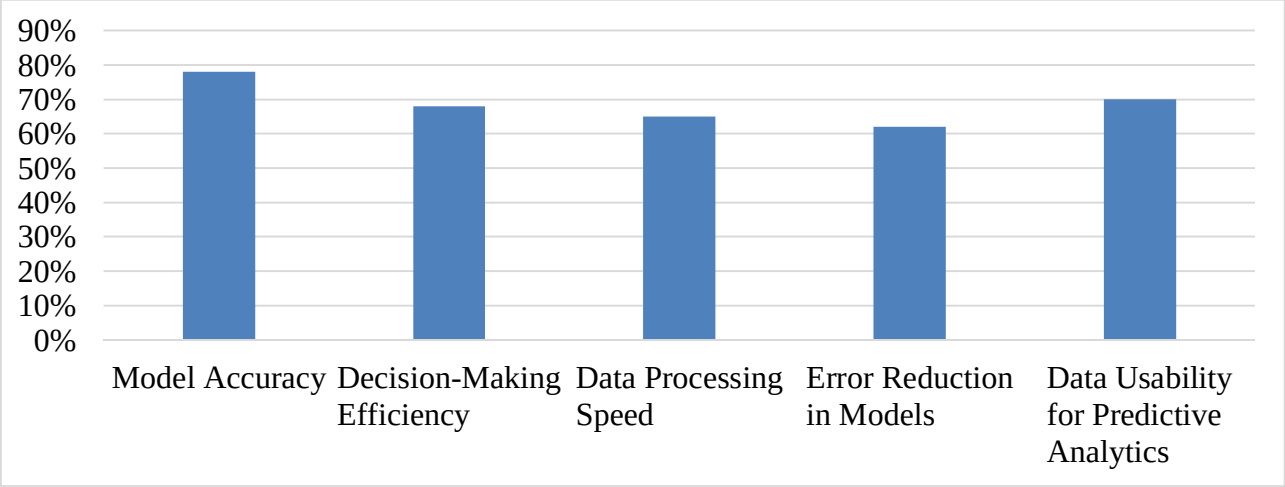


Figure 1: Graphical representation of High Data Quality's Effect on Data Science Results

With 83% of respondents reporting greater compliance with data rules, effective data governance has a significant impact on the outcomes of data science. This is a reflection of the function that governance plays in ensuring that legal standards are adhered to. There was an eighty percent improvement in data security, which highlights the significance that governance plays in securing sensitive information. Furthermore, 74% of respondents reported an increase in the availability and accessibility of data, which suggests that governance helps to streamline data management and guarantees that data is easily accessible for analytics. As a result of governance's ability to allow improved communication and data sharing among teams, 69% of respondents claimed experience with improved collaboration. Last but not least, 66% of respondents reported fewer errors in data pipelines, which demonstrates governance's capacity to preserve data integrity. The overall quality of data, as well as security, compliance, accessibility, and cooperation, are all improved when good governance is implemented in data science projects.

Table 2: Effective Data Governance's Effect on Data Science Results

Aspect of Data Science	Percentage Reporting Improvement (%)
Compliance with Data Regulations	83%
Data Security	80%
Data Availability and Accessibility	74%
Improved Collaboration	69%
Error Reduction in Data Pipelines	66%

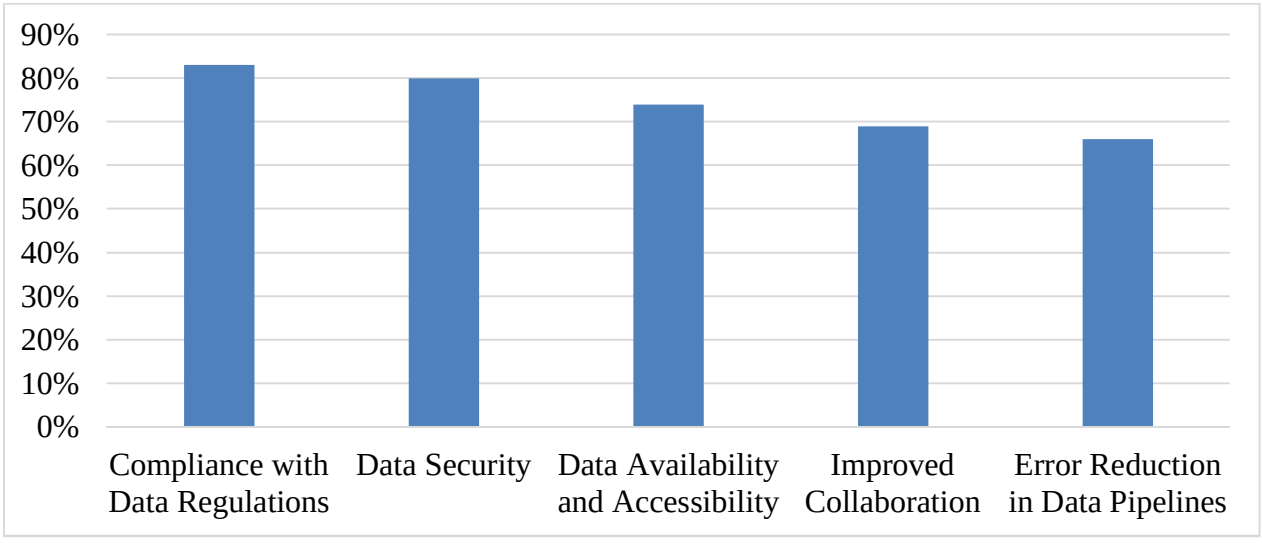


Figure 2: Graphical representation of Effective Data Governance's Effect on Data Science Results

As demonstrated in Table 3 and Figure 3, the combined impact of good data quality and effective governance on the outcomes of data science is significant. There was an increase of 85 percent in the accuracy of the model, which suggests that the combination of high-quality data and robust governance mechanisms results in more trustworthy predictive models. 80% of respondents reported an increase in the efficiency of decision-making, as the combination ensures that accurate and well-governed data supports decisions that are made more quickly and with greater information. When both the quality of the data and the governance of the data are enhanced, the pace at which the data is processed improves by 75%, which reflects smoother processes. It was claimed that there was a 73% reduction in errors in models, which highlights the fact that there are less mistakes when these issues are jointly handled. It is important to note that compliance with data rules had the greatest improvement, which was 88%. This highlights the significant role that governance plays in ensuring that regulatory compliance is maintained alongside high-quality data. Within the context of data science, the combined influence of these elements results in huge improvements across all of the most important parts of the field.

Table 3: The Joint Effect of Good Governance and High Data Quality on Data Science Outcomes

Aspect of Data Science	Percentage Reporting Improvement (%)
Model Accuracy	85%
Decision-Making Efficiency	80%
Data Processing Speed	75%
Error Reduction in Models	73%
Compliance with Data Regulations	88%

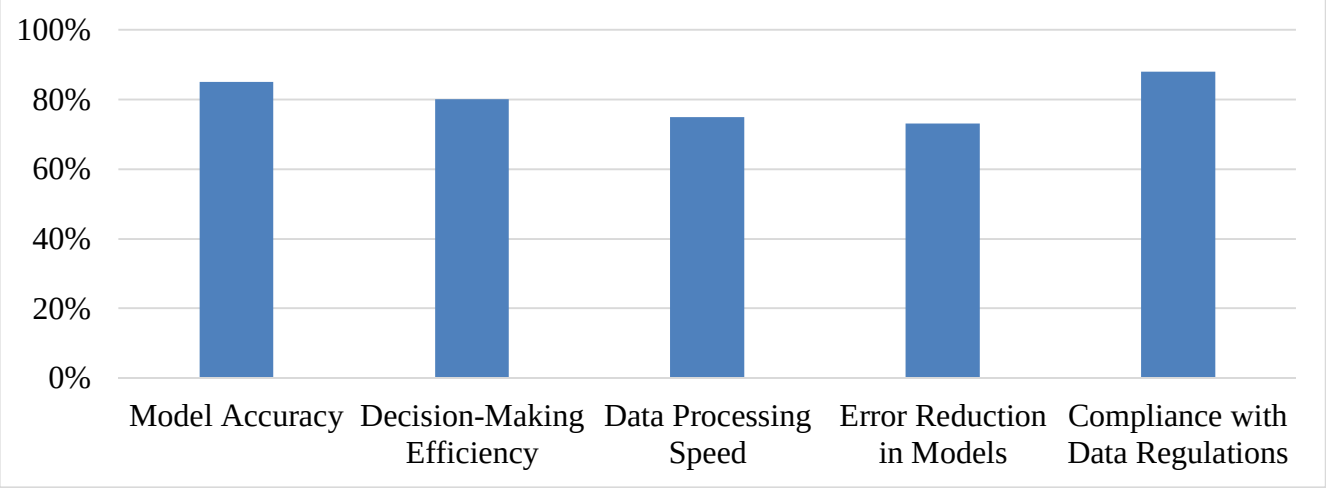


Figure 3: Graphical representation on The Joint Effect of Good Governance and High Data Quality on Data Science Outcomes

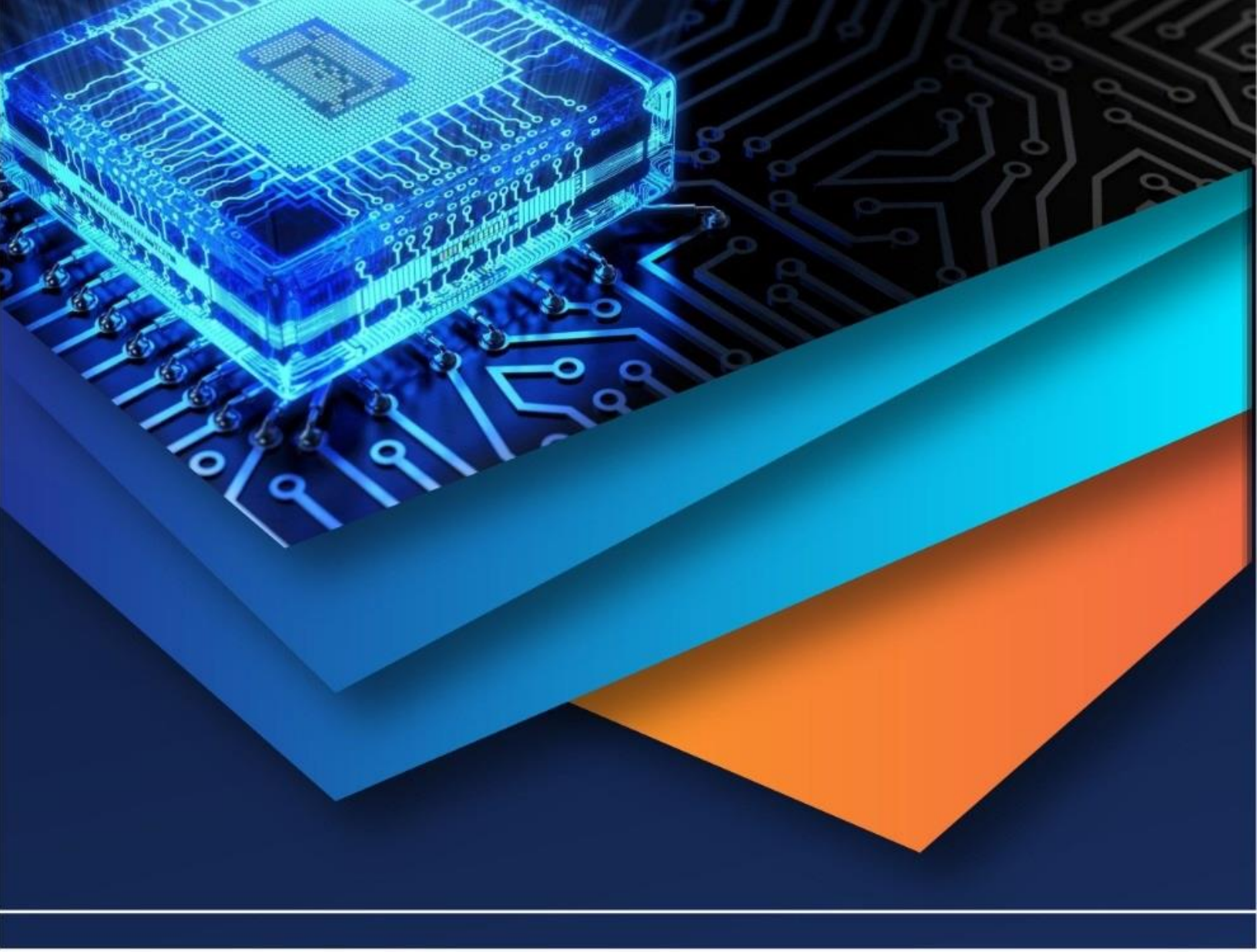
V. CONCLUSION

The findings of this study demonstrate the significant impact that good data quality and efficient data governance have in improving the outputs of data science, which eventually results in enhanced decision-making processes inside businesses. The results of the study reveal that having good data quality substantially improves the accuracy of models, the efficiency with which decisions are made, and the elimination of errors. On the other hand, having strong governance assures compliance with regulations, data security, and accessibility. By prioritizing data quality and governance standards, companies are able to obtain more dependable and trustworthy outcomes. This is demonstrated by the fact that the synergy between these two criteria results in large gains across all critical aspects of data science. Consequently, in order for companies to effectively employ data science, it is necessary for them to make investments in robust data management frameworks and ensure high standards of data quality. These are critical steps toward implementing successful and responsible digital transformation programs.

REFERENCES

[1] Micheli, M., Ponti, M., Craglia, M., & Berti Suman, A. (2020). Emerging models of data governance in the age of datafication. Big Data & Society, 7(2), 2053951720948087.

[2] Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy Artificial Intelligence. Government information quarterly, 37(3), 101493.
[3] Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2021). Data quality of platforms and panels for online behavioral research. Behavior research methods, 1-20.
[4] Yallop, A. C., Gică, O. A., Moisescu, O. I., Coroș, M. M., & Séraphin, H. (2023). The digital traveller: implications for data ethics and data governance in tourism and hospitality. Journal of Consumer Marketing, 40(2), 155-170.
[5] Davidson, E., Wessel, L., Winter, J. S., & Winter, S. (2023). Future directions for scholarship on data governance, digital innovation, and grand challenges. Information and Organization, 33(1), 100454.
[6] Ravikumar, R., Kitana, A., Taamneh, A., Aburayya, A., Shwedeh, F., Salloum, S., & Shaalan, K. (2023). The Impact of Big Data Quality Analytics on Knowledge Management in Healthcare Institutions: Lessons Learned from Big Data's Application within The Healthcare Sector. South Eastern European Journal of Public Health.
[7] Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. Plos one, 18(3), e0279720.
[8] de Hond, A. A., Leeuwenberg, A. M., Hooft, L., Kant, I. M., Nijman, S. W., van Os, H. J., ... & Moons, K. G. (2022). Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ digital medicine, 5(1), 2.
[9] Bag, S., Wood, L. C., Xu, L., Dhamija, P., & Kayikci, Y. (2020). Big data analytics as an operational excellence approach to enhance sustainable supply chain performance. Resources, conservation and recycling, 153, 104559.
[10] Yu, W., Zhao, G., Liu, Q., & Song, Y. (2021). Role of big data analytics capability in developing integrated hospital supply chains and operational flexibility: An organizational information processing theory perspective. Technological Forecasting and Social Change, 163, 120417.
[11] Chen, P. T., Lin, C. L., & Wu, W. N. (2020). Big data management in healthcare: Adoption challenges and implications. International Journal of Information Management, 53, 102078.
[12] Secinaro, S., Calandra, D., Secinaro, A., Muthurangu, V., & Biancone, P. (2021). The role of artificial intelligence in healthcare: a structured literature review. BMC medical informatics and decision making, 21, 1-23.
[13] Barlette, Y., & Baillette, P. (2022). Big data analytics in turbulent contexts: towards organizational change for enhanced agility. Production Planning & Control, 33(2-3), 105-122.
[14] Andronie, M., Lăzăroiu, G., Iatagan, M., Hurloiu, I., & Dijmărescu, I. (2021). Sustainable cyber-physical production systems in big data-driven smart urban economy: a systematic literature review. Sustainability, 13(2), 751.
[15] Yanamala, A. K. Y., & Suryadevara, S. (2023). Advances in Data Protection and Artificial Intelligence: Trends and Challenges. International Journal of Advanced Engineering Technologies and Innovations, 1(01), 294-319.
[16] Gong, Y., & Janssen, M. (2021). Roles and capabilities of enterprise architecture in big data analytics technology adoption and implementation. Journal of theoretical and applied electronic commerce research, 16(1), 37-51.
[17] Kuziemski, M., & Misuraca, G. (2020). AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. Telecommunications policy, 44(6), 101976.
[18] Haendel, M. A., Chute, C. G., Bennett, T. D., Eichmann, D. A., Guinney, J., Kibbe, W. A., ... & Gersing, K. R. (2021). The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. Journal of the American Medical Informatics Association, 28(3), 427-443.
[19] Maroufkhani, P., Tseng, M. L., Iranmanesh, M., Ismail, W. K. W., & Khalid, H. (2020). Big data analytics adoption: Determinants and performances among small to medium-sized enterprises. International journal of information management, 54, 102190.
[20] Yasmin, M., Tatoglu, E., Kilic, H. S., Zaim, S., & Delen, D. (2020). Big data analytics capabilities and firm performance: An integrated MCDM approach. Journal of Business Research, 114, 1-15.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY