

Article

Not peer-reviewed version

The Use of the BEPP-DS Methodology in the Development of Evidence-Based Public Policies (EBPP) in Education Using Data Science (DS)

[Mauricio Feijó Benevides de Magalhães Filho](#) * and [Joaquim Escola](#)

Posted Date: 15 October 2025

doi: 10.20944/preprints202510.1193.v1

Keywords: data science; evidence-based public policy; electronic government; public administration; big data; policy analysis; digital governance; decision support systems



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

The Use of the BEPP-DS Methodology in the Development of Evidence-Based Public Policies (EBPP) in Education Using Data Science (DS)

Mauricio Feijó Benevides de Magalhães Filho ^{1,*} and Joaquim Escola ²

¹ University of Trás-os-Montes and Alto Douro
² University of Porto, University of Trás-os-Montes and Alto Douro
* Correspondence: mauricio.benevides@ufc.br

Abstract

The advancement of technologies and the enormous volume of available data have made Data Science (DS) indispensable for addressing complex social issues. In the public sector, Evidence-Based Public Policies (EBPP) seek to support decisions through data analysis, increasing transparency and governmental effectiveness. However, the adoption of Data Science in public policy faces obstacles such as data quality, the need for collaboration between different areas, and institutional resistance. This article proposes the BEPP-DS methodology, which structures the entire process—from problem identification to policy evaluation—with a focus on transparency, reproducibility, and scalability. The model serves as a reference for governments wishing to use data science to build more effective policies, promoting the use of artificial intelligence, advanced analytics, and citizen engagement in data governance.

Keywords: data science; evidence-based public policy; electronic government; public administration; big data; policy analysis; digital governance; decision support systems

1. Introduction

The exponential growth of data and advances in computational tools have consolidated Data Science (DS) as a key field for tackling complex social challenges (Silva et al., 2022). By facilitating the extraction of conclusions from data, DS has become an essential component in several areas, such as medicine, education, and urban administration (Santos et al., 2023).

Recently, applications of DS in governance have proven to be significant, especially in the formulation and implementation of Evidence-Based Public Policies (EBPP) (Anderson et al., 2005). There is a growing need for governments around the world to make their policies more effective, fair, and transparent, with the full involvement of all relevant stakeholders (Saltelli et al., 2020). Evidence-based approaches enable the systematic use of data in the policy formulation process, ensuring that decisions are grounded in concrete evidence, rather than intuition or tradition.

However, the integration of DS in public policy making presents challenges such as the availability and quality of data, the need for interdisciplinary collaboration, and the application of ethical and systematic methodologies that guide the extraction of knowledge within the context of governmental institutions (MacArthur et al., 2022).

This scenario is particularly relevant for Brazil, where the socioeconomic landscape is marked by regional disparities. Despite these challenges, there is an emerging consensus on the need to structure and formalize the use of DS in government to promote more effective public policies. This work describes a methodology for the construction of EBPP using DS, based on the principles of transparency and non-discrimination.

Given this context, the guiding question of this research is: how can a replicable methodology be structured for the construction of evidence-based public policies, using Data Science principles, applicable in different governmental contexts and capable of promoting more informed and effective decisions?

This article details the developed methodology, enhanced by lessons learned during the implementation of large-scale data platforms, such as the Antonieta de Barros platform (FNDE), among others documented in scientific articles (Sucupira Furtado et al., 2023a) and (Batista et al., 2024).

The remainder of this article is organized as follows: Section 2 presents an overview of the theoretical foundation of Evidence-Based Public Policies and Data Science, highlighting their points of convergence in governance. Section 3 reviews related methodologies and identifies gaps in the existing literature. Section 4 describes the proposed methodology, its core components, and activities in depth. Sections 5, 6, and 7 discuss the lessons learned and future opportunities, especially the integration of Artificial Intelligence in data-driven policy formulation.

2. Theoretical Framework

This section addresses the essential concepts for understanding the intersection between Evidence-Based Public Policies (EBPP) and Data Science in the context of electronic government, especially in Brazil. Understanding these domains helps to recognize the opportunities and challenges of Data Science in governance practices.

2.1. Evidence-Based Public Policies

EBPP are strategies based on empirical data and scientific research, ensuring that decisions are effective and transparent (Anderson et al., 2005). This methodical approach allows governments to systematically address challenges. However, implementation faces challenges such as data quality and institutional resistance. Despite this, the demand for transparency makes EBPP fundamental in modern governance.

2.2. Data Science for Governments

Data Science combines statistical analysis, computational techniques, and domain knowledge to solve complex problems and support decision-making (Van Der Aalst and van der Aalst, 2016). The data science lifecycle includes data collection, cleaning, analysis, interpretation, and dissemination (Rahul and Banyal, 2020). Each phase contributes to the extraction of actionable insights and continuous improvement, which is essential for dynamic environments such as governance.

2.3. Electronic Government in Brazil

Electronic Government (e-Government) is the use of digital technologies to improve the delivery of government services, enhance citizen engagement, and optimize administrative processes (Silcock, 2001). In Brazil, Electronic Government initiatives represent an opportunity to reduce regional disparities and increase the efficiency of the public sector (Musafir, 2018). With more than 200 million inhabitants and a growing internet penetration rate, Brazil has the potential to use digital platforms to expand access to public services and increase transparency. However, the adoption of Electronic Government in Brazil still faces challenges such as deficiencies in rural digital infrastructure, socioeconomic disparities, and bureaucratic resistance.

3. Related Works

This section presents the studies analyzed to understand the models and their respective results provided by Evidence-Based Public Policies and Data Science. Understanding these works allows a better comprehension of the challenges of applying data science in governance practices, especially within Brazil's social contexts.

Sucupira (Sucupira Furtado et al., 2023b) demonstrated how the digital transformation in Ceará integrated technological innovation into the cycles of public policy to address social challenges and promote sustainable development. Tools such as big data platforms and mobile applications enabled data-driven policy formulation, improving efficiency, transparency, and responsiveness. Collaboration among the public, academic, and private sectors strengthened institutional capacity and policy legitimacy. The focus on vulnerable populations, aligned with the Sustainable Development Goals (SDGs), contributed to reducing inequalities through targeted interventions.

Pereira (Pereira et al., 2021) highlighted the role of standardized information structures in creating efficient and accessible policies. Interoperability among government systems improved evidence-based decision-making and digital inclusion by simplifying access to services. Standardized taxonomies enabled better monitoring and evaluation of policies, ensuring adaptive and inclusive governance.

Furtado (Furtado et al., 2023) presented the importance of integrating data and advanced analytics to support the creation of more effective and targeted public policies. By consolidating information on social vulnerabilities—such as income, housing, and nutrition—the system made it possible to identify high-risk populations and prioritize government actions in a well-founded manner. This approach demonstrates how digital tools can strengthen the diagnosis of social problems, offering an empirical basis for formulating policies that are more aligned with the actual needs of the population. A central feature of this work is the ability to personalize policy recommendations based on detailed analyses of specific conditions. Through interactive dashboards, public managers can access tailored solutions to meet particular demands, such as providing financial aid, housing policies, or access to government support programs. This personalization reflects a strategic approach to policy formulation by focusing on practical results and immediate impacts for vulnerable populations, maximizing the efficiency and social impact of government interventions.

Schroer (Schroer et al., 2021) presents CRISP-DM (Cross-Industry Standard Process for Data Mining), highlighting its relevance as a methodological framework capable of guiding the creation of evidence-based public policies. Divided into six phases—business understanding, data understanding, data preparation, modeling, evaluation, and deployment—this model offers a systematic approach to analyze large volumes of data, transform information into useful knowledge, and translate analytical insights into concrete actions. This methodology has strong potential for application in the public sector, where data analysis can enhance the formulation of policies addressing complex social problems. The initial phase of business understanding is essential to align data analysis objectives with public policy priorities. It enables managers to identify critical issues, such as social inequality or low coverage of health services, and to define clear intervention goals. By establishing these connections between data and objectives, the model enables the creation of strategic policies aligned with the real needs of the population. This approach demonstrates how data structuring in analytical projects can improve the diagnosis of problems and direct resources to areas of impact. The data preparation and modeling phases are equally important for the policy formulation process, as they ensure the quality and usability of the analyzed information. By consolidating data from different sources, such as government databases or sectoral statistics, CRISP-DM helps create a comprehensive view of the challenges faced by society. In addition, analytical models developed at this stage allow scenario simulation and the prediction of policy effects before their implementation, reducing uncertainty and optimizing expected results. This is especially relevant in contexts such as health, education, and public safety, where well-planned interventions can generate significant transformations.

The analysis of the studies presented reveals a convergence around the use of digital technologies and structured methodologies to strengthen the process of creating and implementing public policies. However, they also highlight gaps that need to be addressed to ensure comprehensive and sustainable implementation. These studies emphasize both the role of technology in the modernization of public governance and the challenges associated with inter-institutional integration, data privacy, continuous monitoring, and social participation. The use of structured

frameworks demonstrates the importance of defined analytical processes to guide data-driven decision-making. This approach allows the identification of critical problems, the definition of clear goals, and the assessment of the potential impact of policies before implementation. However, inter-institutional integration—necessary for consolidating data and promoting a holistic view of social demands—still faces challenges such as information standardization and the definition of responsibilities among government agencies. Overcoming these challenges is crucial to maximizing the effectiveness of public interventions.

Furthermore, the application of digital transformation in specific contexts reinforces the need to align technology and governance with strategic planning that prioritizes inclusion, sustainability, and data privacy protection. However, the studies leave open questions about how to ensure the financial and operational sustainability of these solutions and how to balance data use with security guarantees and ethical treatment. They also highlight gaps in post-implementation monitoring. Time is needed to assess the long-term impacts of these strategies derived from public policies. These studies emphasize that achieving modern and sustainable public governance requires the integration of technology, ethics, interinstitutional collaboration, and citizen participation. Such integration is essential to ensure efficiency, equity, and significant social impact.

4. The Use of the BEEP-DS Methodology in the Development of the Antonieta de Barros Platform

Based on our experience in governmental Big Data projects, such as the development of the Antonieta de Barros data platform for the National Fund for the Development of Education (FNDE), which integrates data and artificial intelligence tools to improve the design, execution, and monitoring of public policies implemented by the FNDE, it was possible to identify and extract best practices to compose a methodology for Evidence-Based Public Policy Development through Data Science (BEEP-DS). The structuring of the BEEP-DS methodology followed a set of guiding requirements, derived from the analysis of real cases and the limitations observed in previous projects.

The main requirements include: (i) clarity and traceability of decisions; (ii) minimization of rework and data reprocessing; (iii) integration of different professional profiles with defined roles; (iv) feasibility of replication in different institutional contexts; and (v) support for impact assessment.

Drawing on the experiences of the Federal University of Ceará and the resources provided by FNDE, the process of creating data products is structured around six main roles: the requester, the data scientist, the business analyst, the data engineer, the developer, and the data manager. The requester is the FNDE representative responsible for submitting a data product request. This individual presents a strategic question already approved by FNDE, representing a business need previously analyzed before the process begins. Consequently, the submitted question will be structured and ready to be addressed.

The data scientist, responsible for the next phase, is an information technology professional experienced in data analysis. Their role is to assess whether the strategic question meets essential criteria, such as objectivity, well-defined scope, and feasibility within a reasonable timeframe. If necessary, the data scientist collaborates with the requester to refine the question. This professional must monitor the entire process of creating the data product, ensuring that efforts remain aligned with the requester's objectives. Once the question is sufficiently clear and feasible, it is passed to the business analyst.

At this stage, the process aligns with the Business Understanding phase in CRISP-DM, where objectives are identified and success criteria are established. Although predictive analysis is performed by the data scientist, it is essential to understand that decisions regarding the development of data products for public policies require a broader perspective. Therefore, it is proposed that this stage evolve into a collegiate assessment, involving subject matter experts and public management representatives, thus considering not only technical and computational aspects but also the potential

impact of the policy to be strengthened, such as urgency, population coverage, or the mitigation of social risks.

The business analyst identifies the necessary data sources to address the strategic question, leveraging available institutional knowledge. If the data are not accessible or are too costly, the request is returned with feedback; otherwise, it is forwarded (with documentation) to the data engineer, integrating with the Data Understanding phase of CRISP-DM. The data engineer manages the technical infrastructure (connection strings, queries, access tokens). Should obstacles arise (for example, ownership of external data), the issue is escalated to the analyst and data scientist. Successfully obtained data are documented, reflecting the Data Preparation phase of CRISP-DM.

The data scientist then assesses feasibility (strategic suitability, technical feasibility, cost-effectiveness) with the requester. If approved, they jointly define the product (dashboard, chart, etc.), initiating development (Modeling in CRISP-DM). The developer builds the product using a sample dataset (provided by the data manager), while the manager prepares the final dataset. The data scientist validates the product with real data, requesting adjustments if necessary (Evaluation phase). Finally, the requester reviews the product. If approved, the team deploys it, and the data manager catalogs it for use (Deployment in CRISP-DM). If rejected, the data scientist re-evaluates the process.

5. Discussion

The BEPP-DS methodology was developed to address the gaps observed in the application of traditional data science frameworks within the context of public policy. Although CRISP-DM is widely recognized for its six-phase structure—from business understanding to deployment—its direct application in governmental environments can be challenging due to its insufficient emphasis on public sector specificities, such as institutional complexity and the need for transparency and accountability. In comparison, BEPP-DS adapts and expands upon the principles of CRISP-DM by introducing stages tailored to public sector realities. For example, while CRISP-DM focuses on business understanding, BEPP-DS emphasizes strategic issues aligned with political and social priorities. Additionally, BEPP-DS introduces a data governance framework that addresses privacy, ethics, and responsible data usage—areas often overlooked by more generalized frameworks.

Although the methodology has been applied in several projects, such as the development of the Antonieta de Barros platform, a systematic evaluation of its effectiveness and adaptability in different governmental contexts is still required. This includes further case studies, comparative analyses with other methodologies, and performance metrics to demonstrate its benefits and limitations. Moreover, the implementation of BEPP-DS demands significant institutional maturity and technical capacity. Public organizations with less developed structures may face challenges in fully adopting this methodology. Therefore, adaptations and ongoing support are essential to ensure its effectiveness and long-term sustainability.

BEPP-DS represents an advance in the integration of data science into public policy, offering a more contextualized approach sensitive to the needs of the public sector. However, its consolidation as a standard methodology will depend on further validation, adaptation to various institutional realities, and the strengthening of the technical capacities of the organizations involved.

6. Lessons Learned

Based on the BEPP-DS methodology and the use cases examined, a set of lessons learned was systematized. This section discusses these lessons using the following structure: each lesson is presented with a title, a description of the context, the challenge faced, and the actions taken to address it.

- **Identifying Real Problems of Each Institution:** Understanding the specific challenges of each institution is essential for effectively implementing data-driven solutions. In many cases, institutions lack a clear vision of their operational obstacles, resulting in poorly defined goals and inadequate resource allocation. To address this, it is recommended to conduct in-depth

- diagnostics before starting any project. These assessments should focus on understanding institutional workflows, key performance indicators, and stakeholder expectations.
- Collaboratively Building the Strategic Issues to Be Addressed: One of the most significant challenges in formulating evidence-based policies is defining strategic and actionable issues. These questions must be relevant, measurable, and aligned with the institution's objectives. A collaborative approach involving all stakeholders—including policymakers, data scientists, and end users—ensures that the issues address real needs. Facilitated workshops and iterative refinement processes are recommended to align these diverse perspectives.
 - Establishing a Pilot to Rapidly Validate the Methodology: Large-scale projects often face risks associated with untested methodologies. Pilots serve as controlled environments to test assumptions, refine processes, and gather feedback. By implementing a pilot phase, organizations can identify potential bottlenecks or risks. It is advisable to select a small, representative sample for the pilot to maximize its relevance and scalability.
 - Systematizing the Recording of New Strategic Issues: In dynamic environments, new challenges and opportunities continually arise. Without a systematic approach, these new strategic needs may go unnoticed or be poorly documented. Implementing a structured repository or formal process to capture and review new strategic issues ensures that the organization can effectively adapt its policies and strategies over time.
 - Creating Automated Flows to Keep Data Products Updated: Data products often lose relevance if not regularly updated. Many institutions rely on manual processes, which are prone to delays and errors. Automating Extract, Transform, and Load (ETL) processes ensures that data products remain current and accurate. Institutions should invest in robust automation tools and ensure adequate training for their technical teams.
 - Promoting a Data Culture within Institutions: The adoption of data-driven methodologies requires more than technical tools; it demands a cultural shift. Many public institutions encounter resistance to change due to unfamiliarity with data practices. Building a data culture involves continuous education, transparent communication, and the promotion of data-driven decision-making at all organizational levels.
 - Establishing Standards for System Integration: Fragmented systems often lead to inefficiencies and missed opportunities for holistic insights. Integration challenges are particularly acute in multi-agency environments, where data interoperability is essential. The development and application of standards for system integration can address these challenges. Such standards should include data formats, APIs, and security protocols to ensure seamless collaboration.
 - Defining Data Architectures Independent of Business Domains: Rigid architectures tied to specific business domains often limit scalability and adaptability. Designing modular and domain-agnostic data architectures enables institutions to reuse data and infrastructure for multiple applications. This approach facilitates integration with external systems and prepares the organization for evolving future needs.

7. Conclusions

The BEPP-DS approach presented in this article proposes the integration of Data Science into the development of Evidence-Based Public Policies, with an emphasis on transparency, scalability, and efficiency. Inspired by practical experiences, the methodology addresses the inherent obstacles in implementing data-driven strategies in the public sector.

This structure offers clear guidance for each stage of data product creation—from defining priority strategic issues to generating insights that can be applied to governmental decision-making.

The results obtained highlight the importance of collectively defining problems, conducting systematic tests through pilot projects, and institutionalizing data-driven practices. The application of BEPP-DS has proven its usefulness by overcoming recurring obstacles such as data fragmentation, weak organizational data culture, and a lack of automated processes. It has also enabled policy-

making teams to better prioritize and implement more effective interventions, fostering more appropriate responses to relevant social challenges.

For the future, it is suggested to continuously improve the BEPP-DS methodology by incorporating Artificial Intelligence and Machine Learning resources to promote predictive analysis and real-time decision-making. It is also important to create ways to expand social engagement, encouraging the active participation of different stakeholders in the process of collecting, validating, and evaluating policies. Furthermore, it is fundamental to address issues related to ethics and privacy, ensuring that data-driven governance respects the principles of justice, responsibility and inclusion.

References

1. Anderson, L. M., Brownson, R. C., Fullilove, M. T., Teutsch, S. M., Novick, L. F., Fielding, J., & Land, G. H. (2005). Evidence-based public health policy and practice: Promises and limits. *American journal of preventive medicine*, 28(5), 226–230.
2. Batista, É., Andrade, R. M., Santos, I. S., Nogueira, T. P., Oliveira, P. A., Lelli, V., & Oliveira, V. T. (2024). Fortaleza city hall strategic planning based on data analysis and forecasting. *Congresso Ibero-Americano em Engenharia de Software (CIBSE)*, 433–436.
3. Furtado, L. S., Moura, G., Vasconcelos, D. J., Fernandes, G. S., Cruz, L. A., Magalhães, R. P., & Coelho da Silva, T. L. (2023). An analytical citizen relation management system (czrm) for social vulnerability mapping and policy recommendation in brazil. *Decision Support Systems*, 172, 113995. DOI: <https://doi.org/https://doi.org/10.1016/j.dss.2023.113995>.
4. MacArthur, B. D., Dorobantu, C. L., & Margetts, H. Z. (2022). Resilient government requires data science reform. *Nature Human Behaviour*, 6(8), 1035–1037.
5. Musafir, V. E. N. (2018). Brazilian e-government policy and implementation. *International E-Government Development: Policy, Implementation and Best Practice*, 155–186.
6. Pereira, G., Monteiro, I., Vasconcelos, D., Braz, L., & Silva, C. (2021). Classificação taxonômica de categorias de serviços públicos para aplicações digitais. *Anais do IX Workshop de Computação Aplicada em Governo Eletrônico*, 119–130. DOI: <https://doi.org/10.5753/wcge.2021.15982>.
7. Rahul, K., & Banyal, R. K. (2020). Data life cycle management in big data analytics. *Procedia Computer Science*, 73, 364–371.
8. Saltelli, A., Bammer, G., Bruno, I., Charters, E., Di Fiore, M., Didier, E., Nelson Espeland, W., Kay, J., Lo Piano, S., Mayo, D., Pielke Jr, R., Portaluri, T., Porter, T. M., Puy, A., Rafols, I., Ravetz, J. R., Reinert, E., Sarewitz, D., Stark, P. B., ... Vineis, P. (2020). Five ways to ensure that models serve society: A manifesto. *Nature* 582(7813), 482–484. DOI: <https://doi.org/10.1038/d41586-020-01812-9>.
9. Santos, I. S., Oliveira, P. A. M., Oliveira, V. T., Nogueira, T. P., Dantas, A. B. O., Menescal, L. M., Batista, É., & Andrade, R. M. C. (2023). Big Data Fortaleza: Plataforma Inteligente para Políticas Públicas Baseadas em Evidências. *Workshop de Computação Aplicada em Governo Eletrônico (WCGE)*, 200–211. DOI: <https://doi.org/10.5753/wcge.2023.230796>.
10. Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying crisp-dm process model [CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020]. *Procedia Computer Science*, 181, 526–534. DOI: <https://doi.org/https://doi.org/10.1016/j.procs.2021.01.199>.
11. Silcock, R. (2001). What is e-government. *Parliamentary affairs*, 54(1), 88–101.
12. Silva, W. C. P., Macedo, J. A. F. D., & De Queiroz Neto, J. F. (2022). Usando um modelo de classificação para a adequada implantação do patrulhamento policial para o enfrentamento à assaltos a bancos no nordeste do Brasil. *Revista Brasileira de Ciências Policiais*, 13(9), 185–205. DOI: <https://doi.org/10.31412/rbcp.v13i9.845>.
13. Sucupira Furtado, L., Da Silva, T. L. C., Ferreira, M. G. F., De Macedo, J. A. F., & De Melo Lima Cavalcanti Moreira, J. K. (2023a). A framework for Digital Transformation towards Smart Governance: Using big data

- tools to target SDGs in Ceará, Brazil. *Journal of Urban Management*, 12(1), 74–87. DOI: <https://doi.org/10.1016/j.jum.2023.01.003>.
14. Sucupira Furtado, L., da Silva, T. L. C., Ferreira, M. G. F., de Macedo, J. A. F., & de Melo Lima Cavalcanti Moreira, J. K. (2023b). A framework for digital transformation towards smart governance: Using big data tools to target sdgs in ceará, brazil [Digital Technologies in Urban Planning and Urban Management]. *Journal of Urban Management*, 12(1), 74–87. DOI: <https://doi.org/https://doi.org/10.1016/j.jum.2023.01.003>.
 15. Van Der Aalst, W., & van der Aalst, W. (2016). *Data science in action*. Springer.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.