MRC: A Hybrid Relevance and Correlation Metric for Indicator Selection with Application to SINASC Data

Daniel de Amaral da Silva[1], José Luciano Martins Neto[2], Adilio J. Freitas[3], Antonio Rafael Braga[4], Danielo G. Gomes[1]

[1]Programa de Pós-Graduação em Engenharia de Teleinformática,
Grupo de Redes de Computadores, Engenharia de Software e Sistemas (GREat)
Centro de Tecnologia, Universidade Federal do Ceará, Fortaleza - CE

[2]Bacharelado em Ciência da Computação
Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE), Iguatu - CE

[3]Bacharelado em Ciência de Dados
Centro de Ciências, Universidade Federal do Ceará, Fortaleza - CE

[4]Redes de Computadores, Campus Quixadá,
Universidade Federal do Ceará, Quixadá - CE

{danielamaral,adiliojfreitas}@alu.ufc.br, {rafaelbraga,danielo}@ufc.br,
luciano.neto03@aluno.ifce.edu.br

Abstract. *Variable selection is a critical step in data analysis and modeling, especially in domains with high-dimensional datasets such as healthcare. Addressing this challenge is particularly crucial in digital governance, where robust indicator prioritization underpins evidence-based policymaking. Traditional approaches often rely solely on statistical correlation or feature importance derived from predictive models, potentially overlooking semantic relevance. In this paper, we introduce the Metric of Relevance and Correlation (MRC), a novel hybrid approach that combines semantic relevance, statistical correlation, and predictive impact to identify the most relevant variables for a target outcome. We apply MRC to the Brazilian Live Birth Information System (SINASC) data to identify key factors associated with the 5-minute APGAR score, an important indicator of newborn health. Our results show that MRC provides a more stable variable ranking compared to traditional correlation-based and random forest-based methods, with superior performance across multiple stability metrics. MRC successfully identifies clinically relevant variables while discovering non-obvious relationships that traditional methods might miss. This approach offers a more comprehensive framework for variable selection, applicable across various domains requiring robust feature prioritization, and is particularly valuable for digital government initiatives in aiding the formulation of more effective public policies.*

1. Introduction

Digital governance increasingly demands evidence-based decision-making and optimal resource allocation while maintaining administrative transparency [Yang et al. 2024]. Government agencies collect vast and diverse datasets, including administrative records,

census surveys, and real-time data from sensors and digital platforms. The integration of these sources is essential for providing policymakers with a comprehensive understanding of social challenges and directing effective interventions [Parimala et al. 2017]. This is a cornerstone for advancing digital governance and optimizing public resource allocation.

A critical challenge in governmental data analysis is identifying which indicators truly matter for specific outcomes. With hundreds of potential variables across interconnected systems, prioritizing the most relevant factors for policy intervention becomes essential [Lynn et al. 2000]. Traditional data analysis approaches often prove insufficient as the management of relevant data sources in a qualitative way necessitates new methods beyond conventional approaches [European Commission 2016], failing to incorporate the rich contextual knowledge embedded in government data dictionaries and metadata, potentially missing semantic connections important for policy formulation.

Feature selection plays a pivotal role in data analysis, particularly in high-dimensional government datasets where identifying the most relevant variables is essential for both model performance and policy interpretability [Guyon and Elisseeff 2003]. While traditional approaches (filter, wrapper, and embedded methods [Saeys et al. 2007]) focus primarily on statistical relationships, they often neglect semantic relevance crucial for governance contexts.

In digital government, advanced analytical methods like NLP and Machine Learning can transform data into valuable policy insights [Alexopoulos et al. 2019], but their effectiveness depends critically on identifying truly relevant indicators from the multitude of variables collected across government systems. Recent developments have addressed these limitations through hybrid approaches, such as feature selection methods considering interaction characteristics in Neighborhood Rough Sets [Wan et al. 2021] and hybrid feature-selection models to analyze regional impacts of community features on COVID-19 case fatality rates [Moosazadeh et al. 2022].

To address these challenges, we propose the Metric of Relevance and Correlation (MRC), a novel hybrid indicator that integrates three complementary dimensions of variable relevance: semantic relevance through natural language processing of variable descriptions, statistical correlation using adaptive methods based on variable types, and predictive impact through evaluation of variables' contributions to outcome prediction. We apply MRC to Brazil's SINASC dataset to identify factors relevant to newborn health outcomes, demonstrating how this approach can directly inform public health policy decisions. Our results show that MRC produces more stable variable rankings compared to traditional approaches, identifying both expected clinical factors and revealing potential geographic disparities in healthcare quality that warrant policy attention.

The MRC framework bridges the gap between purely statistical methods and domain expertise, supporting evidence-based policy formulation in complex domains where understanding the relationships between variables and outcomes requires both statistical rigor and contextual knowledge. By identifying the most relevant variables with greater stability and incorporating domain expertise systematically, MRC helps government agencies prioritize focus areas for intervention, optimize resource allocation, and design more effective policies based on comprehensive data insights.

Table 1 presents the main variables from the SINASC dataset referenced through-

out this paper. To improve readability, we use the standardized abbreviations as defined in the official data dictionary instead of full variable names.

Table 1. Main SINASC Variables Referenced in This Study

Abbreviation	Definition
APGAR1	APGAR score at first minute
APGAR5	APGAR score at fifth minute
CODESTAB	Health facility code
CODMUNNASC	Municipality code of birth occurrence
CODMUNNATU	Municipality code of mother's birthplace
CODMUNRES	Municipality code of mother's residence
CONTADOR	Record counter
DTNASCMAE	Mother's date of birth
DTRECEBIM	Date of receipt at central level
GESTACAO	Pregnancy duration in weeks
HORANASC	Time of birth
NUMEROLOTE	Batch number of records
PESO	Birth weight in grams
SEMAGESTAC	Number of gestation weeks
STDNEPIDEM	Birth certificate status in epidemiological investigation
TPROBSON	Robson classification
MUNRESAREA	Area of mother's municipality of residence
MUNRESLON	Longitude of mother's municipality of residence
MUNRESNOME	Name of mother's municipality of residence

2. Background

The Metric of Relevance and Correlation (MRC) combines three complementary components to evaluate the relevance of variables to a target variable. Each component addresses a different dimension of relevance, providing a comprehensive assessment framework.

$$\text{MRC}_i = \alpha \cdot \text{IRS}_i + \beta \cdot \text{CCE}_i + \gamma \cdot \text{FIC}_i \, , \tag{1}$$

where α, β, and γ are weights assigned to each component, with $\alpha + \beta + \gamma = 1$. The pseudocode for the MRC algorithm is presented in Algorithm 1.

The MRC score represents a comprehensive measure of variable relevance that integrates semantic similarity, statistical correlation, and predictive power. Scores range from 0 to 1, with higher values indicating stronger relevance to the target variable across multiple dimensions. Empowering the users to make context-appropriate interpretations.

2.1. Semantic Relevance Index (IRS)

The Semantic Relevance Index (IRS) measures the semantic similarity between variable descriptions using Natural Language Processing. This component captures the conceptual relevance of each variable to the target variable based on their textual descriptions, which is valuable when working with datasets that include data dictionaries or metadata.

Algorithm 1 MRC Calculation Algorithm

procedure CALCULATEMRC(data, descriptions, target, weights)
    Initialize component weights (α, β, γ)
    Extract variable information from data
    Obtain target variable description
    Calculate Semantic Relevance Index (IRS)
    Calculate Statistical Correlation Coefficient (CCE)
    Calculate Cross-Impact Factor (FIC)
    Compute MRC scores as weighted sum: $\alpha \cdot \text{IRS} + \beta \cdot \text{CCE} + \gamma \cdot \text{FIC}$
    Rank variables by MRC score
    return ranked variables with scores
end procedure

The IRS calculation process involves several key steps. First, a pretrained language model is loaded to generate vector embeddings for the textual descriptions of all variables in the dataset. These embedding models, typically based on transformer architectures, convert text into high-dimensional vector representations that capture semantic meaning and contextual relationships. Once embeddings are generated for all variable descriptions, including the target variable, cosine similarity is calculated between each variable's embedding and the target variable's embedding. This similarity measure quantifies the semantic closeness between concepts [Rahutomo et al. 2012], resulting in scores ranging from 0 (completely unrelated) to 1 (identical meaning).

2.2. Statistical Correlation Coefficient (CCE)

The Statistical Correlation Coefficient (CCE) quantifies the direct statistical relationship between each variable and the target variable. Unlike traditional correlation-based approaches that apply a single correlation method to all variables, the CCE component adaptively selects the most appropriate correlation method based on the types of variables being compared.

The CCE calculation begins by identifying the statistical type of each variable (numerical continuous, numerical discrete, categorical nominal, or categorical ordinal) and the type of the target variable. Based on this information, along with distribution characteristics and sample size considerations [Schober et al. 2018, de Winter et al. 2016], an appropriate correlation method is selected for each variable-target pair. For continuous numerical variables with normal distributions, Pearson's correlation coefficient may be used, while Spearman's rank correlation is more suitable for non-normally distributed data [de Winter et al. 2016]. Kendall's tau might be preferred for small samples or when many tied ranks exist [Xu et al. 2013]. For categorical variables, measures like Cramér's V or the contingency coefficient are appropriate, while mixed numerical-categorical pairs might utilize the eta coefficient [Kennedy 1970]. After calculating the correlation using the selected method, the absolute value is recorded to capture both positive and negative relationships equally.

2.3. Cross-Impact Factor (FIC)

The Cross-Impact Factor (FIC) evaluates the individual predictive power of each variable with respect to the target variable. This component assesses how well each variable, on its

own, can predict the target outcome, capturing the direct predictive relationship without the influence of other variables.

The FIC calculation process begins by selecting an appropriate predictive model based on the target variable type (e.g., linear regression for continuous targets, logistic regression for binary targets). After preprocessing and standardizing the data, for each variable in the dataset, a separate model is trained using only that variable as a predictor for the target variable. The performance of each single-variable model is evaluated using appropriate metrics (e.g., R^2 for regression, accuracy for classification). These individual performance scores are then normalized to a 0-1 scale to make them comparable with the other MRC components.

The FIC provides a complementary perspective to the IRS and CCE by focusing on the direct predictive capability of each variable in isolation. This approach helps identify variables that maintain strong predictive power independently.

3. Materials and Methods

3.1. Dataset

We utilized data from Brazil's SINASC (Live Birth Information System) [DATASUS 2021], which collects comprehensive information on all births nationwide. For this study, we focused on the most recent available data (2023) obtained through the microdatasus package in R. The initial dataset contained 111,091 rows and 75 variables. We performed data preprocessing by filtering columns to include only those with a maximum of 5% missing values, which reduced the dataset to 53 variables. Subsequently, we removed rows with missing values, resulting in a final dataset of 94,765 rows and 53 variables.

For this study, we selected the 5-minute APGAR score as our target variable, a standardized measure developed by Virginia Apgar to assess newborn health [Apgar 1953]. The APGAR score evaluates five components (Appearance, Pulse, Grimace, Activity, and Respiration) on a scale of 0-2 each, resulting in a total score from 0-10, with higher scores indicating better newborn health status.

3.2. Variable Types and Descriptions

Each variable in the SINASC dataset was categorized into one of four types to enable appropriate correlation analysis:

- Numerical Continuous: Variables with continuous values (e.g., mother's age, birth weight)
- Numerical Discrete: Variables with discrete numerical values (e.g., number of prenatal visits)
- Categorical Nominal: Variables with unordered categories (e.g., municipality of birth)
- Categorical Ordinal: Variables with ordered categories (e.g., education level)

Variable descriptions were obtained from the official DATASUS data dictionary[1], providing standardized definitions for each field in the dataset. These descriptions were used for the semantic component of the MRC analysis.

[1] https://datasus.saude.gov.br/transferencia-de-arquivos/

3.3. Implementation of MRC

We implemented the MRC metric following the framework outlined in the section 2 (background). For the semantic component (IRS), we utilized the BERTimbau model [Souza et al. 2020], a Portuguese language variant of BERT pretrained specifically for Brazilian Portuguese. This model was chosen due to its state-of-the-art performance on Brazilian Portuguese language tasks and its ability to effectively capture semantic relationships between healthcare terminology. Using BERTimbau, we generated vector embeddings for each variable description from the data dictionary and calculated cosine similarity between each variable's description embedding and the APGAR5 description embedding.

The statistical component (CCE) was implemented following the adaptive approach described in the section 2, where different correlation methods were applied based on the statistical properties of each variable-target pair. Similarly, for the Cross-Impact Factor (FIC), we trained individual predictive models for each variable as described in the section 2, using Linear Regression for the numerical APGAR5 target and calculating R^2 as the performance metric.

The component weights were set to $\alpha = 0.4$ for IRS, $\beta = 0.3$ for CCE, and $\gamma = 0.3$ for FIC based on preliminary testing. These weights reflect a slightly higher emphasis on semantic relevance while maintaining balance across all three dimensions of variable importance. A detailed sensitivity analysis of these weights is considered a direction for future work (see Section 5.2).

3.4. Comparison Methods

We compared MRC rankings with two traditional feature selection approaches. The first was Univariate Correlation, which ranked variables based solely on their correlation coefficients with APGAR5. The second was Random Forest Importance, where rankings were derived from feature importance scores in a Random Forest model trained to predict APGAR5. These methods were selected to represent widely used statistical and machine learning approaches to variable ranking, providing a meaningful comparison with our hybrid methodology.

3.5. Stability Analysis

To assess the robustness of the variable rankings, we performed bootstrap resampling with 100 iterations. For each bootstrap iteration, we applied all three ranking methods (MRC, Correlation, Random Forest) and recorded the top 8 variables selected by each approach. We then calculated four stability metrics to comprehensively assess different aspects of ranking consistency: Position Stability (average variation in ranking positions), Jaccard Stability (average similarity between selected variable sets), Kuncheva Stability Index (measure accounting for random chance in selection), and Spearman Stability (rank correlation between iterations). These metrics together provide a robust evaluation of how consistently each method identifies and ranks important variables across different data samples.

4. Results

4.1. MRC Rankings and Component Contributions

Table 2 presents the top 10 variables ranked by the MRC metric, along with their individual component scores. The 1-minute APGAR score (APGAR1) emerged as the most relevant variable with the highest MRC score (0.691), followed by mother's birth date (DTNASCMAE) with a score of 0.498.

Table 2. Top 10 Variables Ranked by MRC Score

Variable	MRC	IRS	CCE	FIC
APGAR1	0.691	0.958	0.563	0.463
DTNASCMAE	0.498	0.438	0.655	0.423
CODESTAB	0.332	0.466	0.357	0.127
CODMUNNASC	0.325	0.485	0.329	0.108
CODMUNNATU	0.301	0.483	0.281	0.079
CODMUNRES	0.276	0.419	0.282	0.079
MUNRESNOME	0.266	0.394	0.282	0.079
PESO	0.263	0.577	0.069	0.039
DTRECEBIM	0.263	0.458	0.218	0.048
TPROBSON	0.263	0.570	0.088	0.027

Analyzing the component contributions in Table 2 reveals interesting patterns in how different components affect the final MRC score. APGAR1 shows high scores across all three components, with particularly strong semantic relevance (IRS=0.958), reflecting its close conceptual relationship with APGAR5. In contrast, DTNASCMAE (mother's birth date) has its strongest contribution from statistical correlation (CCE=0.655), suggesting a significant empirical relationship despite lower semantic similarity. Geographic variables (CODESTAB, CODMUNNASC, CODMUNNATU, CODMUNRES) show moderate scores across components, while several variables like PESO (birth weight) and TPROBSON (Robson classification) have high semantic relevance but relatively low statistical correlation and predictive impact. The distribution of scores shows a gradual decline rather than a sharp threshold, reinforcing the value of presenting the full spectrum of MRC scores for flexible interpretation.

4.2. Comparison with Traditional Methods

Table 3 compares the top 8 variables identified by each method. While there is some overlap, particularly with APGAR1 being universally recognized as the most important variable, each method identifies different sets of variables as most relevant. The MRC method places greater emphasis on geographic and healthcare facility variables (CODESTAB, CODMUNNASC, CODMUNNATU, CODMUNRES), which are less prominent in the other approaches. Correlation-based ranking highlights variables related to pregnancy characteristics (GESTACAO, SEMAGESTAC) and administrative aspects (NUMEROLOTE), while Random Forest emphasizes birth details (PESO, HORANASC) and processing information (CONTADOR).

These differences in variable prioritization highlight how each method provides a unique perspective on variable importance, with MRC offering a more balanced approach that incorporates semantic knowledge alongside statistical relationships.

Table 3. Comparison of Top 8 Variables by Different Methods

Rank	MRC	Correlation	Random Forest
1	APGAR1	APGAR1	APGAR1
2	DTNASCMAE	NUMEROLOTE	CODMUNNASC
3	CODESTAB	MUNRESLON	PESO
4	CODMUNNASC	GESTACAO	CONTADOR
5	CODMUNNATU	MUNRESAREA	DTNASCMAE
6	CODMUNRES	SEMAGESTAC	HORANASC
7	MUNRESNOME	STDNEPIDEM	MUNRESLON
8	PESO	TPROBSON	SEMAGESTAC

4.3. Stability Analysis Results

A critical advantage of the MRC method is its superior stability compared to traditional approaches. Figure 1 presents the results of our stability analysis across 100 bootstrap iterations. As shown in the bar chart, MRC consistently outperforms other methods across all stability metrics. MRC achieved perfect Jaccard and Kuncheva stability (1.000), indicating that it consistently selected the same set of variables across bootstrap samples. It also demonstrated the lowest position variation (corresponding to the highest 1/Position value of 1.233) and highest Spearman stability (0.994), confirming that both the selected variables and their relative rankings remained consistent across different data samples.

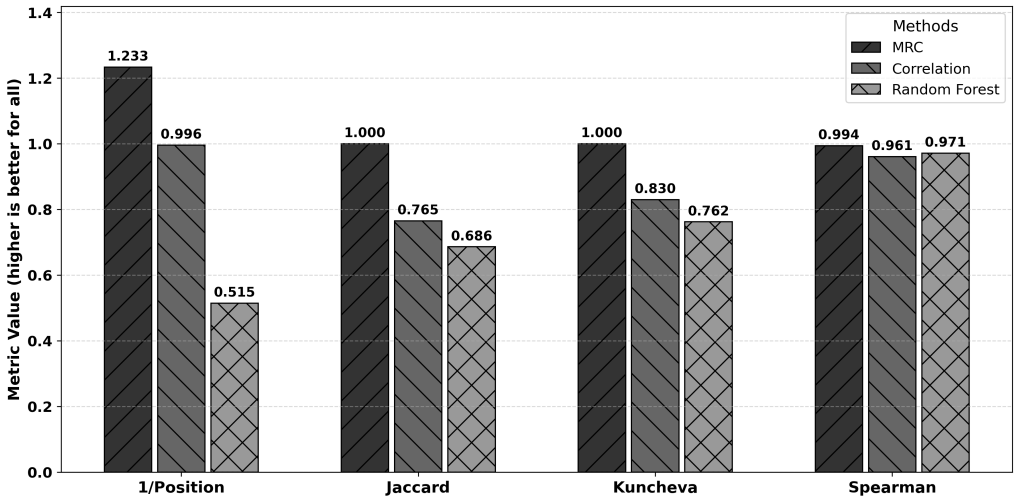

Figure 1. Stability comparison across different variable selection methods. Position is transformed to 1/Position for consistent interpretation with other metrics.

In contrast, the correlation-based method showed moderate stability (Jaccard=0.765, Kuncheva=0.830), while Random Forest exhibited the lowest stability across metrics (Jaccard=0.686, Kuncheva=0.762) with a substantially lower 1/Position value (0.515, corresponding to a position variation of 1.943) compared to MRC. While all three methods maintained relatively high Spearman rank correlations, the gap in Jaccard and Kuncheva metrics clearly demonstrates MRC's superior consistency in selecting the same variables across different data samples, a critical advantage for reliable biomarker identification.

4.4. Variable Categories and Clinical Relevance

Analyzing the variables identified by MRC reveals a diverse set of factors associated with newborn health as measured by the APGAR5 score. The strong association between APGAR1 and APGAR5 aligns with clinical expectations, as the 1-minute score often predicts the 5-minute score. The identification of maternal birth date (likely a proxy for maternal age) and birth weight as relevant factors is also consistent with clinical knowledge about risk factors for neonatal outcomes.

The prominence of geographic and healthcare facility variables suggests potential disparities in healthcare quality and access that may influence newborn outcomes across different regions and facilities in Brazil. This finding is particularly valuable from a public health perspective, as it highlights systemic factors that might not be captured by traditional clinical variables alone.

The MRC approach successfully identified a mix of direct clinical indicators (APGAR1, PESO), maternal characteristics (DTNASCMAE), and system-level factors (healthcare facility and geographic identifiers), providing a more comprehensive view of the determinants of newborn health outcomes than methods focusing solely on statistical relationships.

5. Discussion

5.1. Key Findings and Government Policy Implications

The MRC methodology offers actionable insights for digital governance, particularly concerning factors associated with 5-minute APGAR scores in Brazilian newborns. The strong relationship between APGAR1 and APGAR5 aligns with clinical expectations, while the prominence of maternal birth date (likely a proxy for maternal age) and birth weight reinforces known risk factors for neonatal outcomes.

Particularly significant for public policy is the identification of geographic variables (municipality codes) and healthcare facility identifiers among the top-ranked factors. This suggests potential disparities in healthcare quality and access across different regions of Brazil—a critical finding that warrants attention from policymakers. Such geographic variations could indicate inequitable distribution of healthcare resources, differences in medical practices, or socioeconomic factors affecting maternal and neonatal health.

From a governmental perspective, these findings suggest three key areas for policy intervention:

1. **Healthcare Infrastructure Development:** Targeted improvement of facilities in underperforming regions
2. **Quality Standardization:** Implementation of consistent care protocols across healthcare facilities
3. **Maternal Health Programs:** Enhanced prenatal care especially for high-risk age groups

The ability to identify such specific policy-relevant factors demonstrates MRC's value as a tool for evidence-based governance and resource allocation.

5.2. Methodological Advantages and Limitations

The differences in variable rankings between MRC, correlation-based, and random forest-based methods highlight the unique perspective our approach brings to variable selection. While correlation-based methods favor direct statistical relationships and random forest captures non-linear relationships (albeit with less stability), MRC provides a balanced view incorporating semantic, statistical, and predictive dimensions.

MRC's superior stability across all four metrics (Position, Jaccard, Kuncheva, and Spearman) is particularly valuable in governmental contexts where consistent, reliable variable selection is essential for developing sustainable policies. The perfect Jaccard and Kuncheva stability indices (1.000) indicate that MRC consistently identified the same set of highly-ranked variables across bootstrap samples—a significant advantage when policy decisions must be defended with robust evidence.

The approach does have limitations that should be acknowledged. Semantic relevance depends on the quality of variable descriptions, which may vary in government datasets. The component weights (α, β, γ) may require adjustment across different domains, and the current implementation uses relatively simple predictive models. Additionally, our analysis used a sample of the full SINASC dataset, potentially affecting the ranking of variables in the case of rare outcomes. Specifically, a comprehensive sensitivity analysis for the component weights (α, β, γ) was beyond the scope of this initial study but is a relevant next step.

While our current implementation intentionally uses Linear Regression for FIC to ensure a direct and interpretable assessment of each variable's isolated predictive power, future work could also explore the impact of different machine learning algorithms for the FIC component such as Random Forest, which offers its own variable importance measures.

5.3. Broader Applications in Digital Government

While we demonstrated MRC on Brazilian healthcare data, the methodology has broader applications across digital government initiatives. Government agencies maintain numerous datasets with extensive metadata and data dictionaries, making them ideal candidates for the MRC approach. Potential applications include:

- Identifying key variables in educational outcomes for targeted intervention programs
- Determining the most relevant factors in public transportation efficiency
- Prioritizing indicators for monitoring environmental regulations compliance
- Selecting key metrics for public service performance dashboards

By providing a balanced, stable approach to variable selection, MRC can help bridge the gap between data analysis and policy implementation—a crucial consideration for translating research findings into effective governance.

6. Conclusion

Here we present the Metric of Relevance and Correlation (MRC), a novel hybrid approach for variable selection that combines semantic relevance, statistical correlation, and predictive impact. When applied to Brazil's SINASC dataset for identifying factors associated

with 5-minute APGAR scores, MRC demonstrated superior stability compared to traditional variable selection methods, consistently ranking the same top variables across bootstrap samples.

Our approach successfully identified both clinically relevant variables and potential geographic disparities in healthcare quality that may contribute to variations in newborn outcomes. By presenting comprehensive MRC scores, the method provides a more comprehensive framework for variable selection that aligns better with domain knowledge while producing more reliable results.

The key contributions of this paper include a methodological framework that bridges analytical approaches and domain expertise, an adaptive correlation component that selects appropriate measures based on variable types, empirical demonstration of superior stability in variable selection, and identification of policy-relevant factors in the Brazilian public health system.

Future development of the MRC approach should focus on optimizing component weights for different domains, incorporating more sophisticated predictive models, exploring applications across diverse governmental datasets, and developing user-friendly implementations to facilitate adoption by policy analysts. Additionally, the geographic and facility-related findings warrant deeper investigation to understand specific mechanisms driving regional variations in health outcomes. By providing a more stable and comprehensive approach to variable selection that respects the complexity of interpretation, MRC offers valuable support for both researchers and policymakers working to identify and address key factors affecting public services, ultimately contributing to more evidence-based governance and improved outcomes for citizens.

Acknowledgements

This study was financed in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* - Brazil (CAPES) - Finance Code 001. Danielo G. Gomes gratefully acknowledges support from the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq) through a research productivity grant (process 311845/2022-3), and from the *Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico* (FUNCAP) for funding the Chief Scientist for Digital Transformation project in the state of Ceará (process 06681109/2023).

References

Alexopoulos, C., Lachana, Z., Androutsopoulou, A., Diamantopoulou, V., Charalabidis, Y., and Loutsaris, M. A. (2019). How machine learning is changing e-government. In *Proceedings of the 12th International Conference on Theory and Practice of Electronic Governance*, ICEGOV '19, page 354–363, New York, NY, USA. Association for Computing Machinery.

Apgar, V. (1953). A proposal for a new method of evaluation of the newborn infant. *Anesthesia & Analgesia*, 32(1):260–267.

DATASUS (2021). SINASC - sistema de informações sobre nascidos vivos. *Ministério da Saúde*.

de Winter, J. C., Gosling, S. D., and Potter, J. (2016). Choosing between pearson and spearman correlation coefficients to assess correlations in the context of exposure assessment comparison exercises. *Journal of Exposure Science & Environmental Epidemiology*, 26(5):530–530.

European Commission (2016). Big data analytics for policy making report. Technical report, European Commission, Interoperable Europe. Level of specialisation: Intermediate. Published under the Interoperability Solutions for European Public Administrations programme (2016-2020).

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182.

Kennedy, J. J. (1970). The eta coefficient in complex anova designs. *Educational and Psychological Measurement*, 30(4):885–889.

Lynn, Laurence E., J., Heinrich, C. J., and Hill, C. J. (2000). Studying governance and public management: Challenges and prospects. *Journal of Public Administration Research and Theory*, 10(2):233–262.

Moosazadeh, M., Ifaei, P., Tayerani Charmchi, A. S., Asadi, S., and Yoo, C. (2022). A machine learning-driven spatio-temporal vulnerability appraisal based on socio-economic data for covid-19 impact prevention in the u.s. counties. *Sustainable Cities and Society*, 83:103990.

Parimala, K., Rajkumar, G., Ruba, A., and Vijayalakshmi, S. (2017). Challenges and opportunities with big data. *International Journal of Scientific Research in Computer Science and Engineering*, 5(5):16–20.

Rahutomo, F., Kitasuka, T., and Aritsugi, M. (2012). Semantic cosine similarity. In *The 7th International Student Conference on Advanced Science and Technology (ICAST)*, Seoul, South Korea.

Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517.

Schober, P., Boer, C., and Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23*.

Wan, J., Chen, H., Yuan, Z., Li, T., Yang, X., and Sang, B. (2021). A novel hybrid feature selection method considering feature interaction in neighborhood rough set. *Knowledge-Based Systems*, 227:107167.

Xu, W., Hou, Y., Hung, Y., and Zou, Y. (2013). A comparative analysis of spearman's rho and kendall's tau in normal and contaminated normal models. *Signal Processing*, 93(1):261–276.

Yang, C., Gu, M., and Albitar, K. (2024). Government in the digital age: Exploring the impact of digital transformation on governmental efficiency. *Technological Forecasting and Social Change*, 208:123722.