Modelling Investment Programs with Machine Learning and Data Mining: Descriptive and Predictive Models of Healthcare State Programs in Russia

Aleksandr Gerasimov, Egor Trofimov
National Research University
Higher School of Economics,
The All-Russian State University of Justice
Saint-Petersburg, Russia
atgerasimov@edu.hse.ru
diterihs@mail.ru

Georgy Kopanitsa, Oleg Metsker
ITMO University
Saint-Petersburg, Russia
georgy.kopanitsa@gmail.com
olegmetsker@gmail.com

Abstract—This paper describes the results of data research and the development of models based on machine learning for forecasting and describing factors affecting the achievement of federal projects on the example of a project to improve the quality of life in Russia. We searched for the best predictors to forecast an infant mortality as one of the direct indicators of a healthcare system quality. The dataset included the social and economical indexes of Russia. We performed a Gradient Boost with 0.98 R squared. We performed a correlation analysis using F-Measure to find the predictors with the highest predictive power. During the research work, it was possible to identify factors affecting the indicators of the national health project. It is worth noting that the indicators of national projects are influenced by the level of education of the head of the region and work experience. The combination of a large territory, a high proportion of the urban population and economic activity are three important factors that require the modernization of the system of public administration. However, in Russian conditions, when public administration is highly unified, such a transformation practically does not take place. This work certainly contributes to the development of methods of applying machine learning to analyze complex performance indicators of public administration. The conclusion is made about the possibilities of using the obtained model to supplement the program-objective methods used in government programs with more fine and precise regulatory tools to achieve greater efficiency by optimizing the law.

I. INTRODUCTION

A large amount of current literature on the use of IT and big data analysis in public administration is mainly focused on the investigation of the e-government projects implementation, open data, and e-public services for citizens [1]. New technologies of data analysis and processing are seen as additional tools for implementing existing information practices of public administration [2], economic management [3] or interaction between the government and citizens [4] rather than as major decision-making factors in planning, development, and implementation of public policies [5]. The application of big data in policy making has a much greater potential to achieve practical outcomes. In existing scientific practice there are works describing in detail the methods of building mechanisms for collecting, processing, and presenting information for more effective management of society. These may be the first one presents the idea that the main application of big data in management is to create a qualitatively new architecture of maintaining state databases, which would integrate not only official statistics and departmental reporting but would also have a certain set of "network indicators" for evaluating this or that area of society [6]. As an example of scientific work in this direction Model-driven data acquisition in sensor networks. In Proceedings of the Thirtieth international conference on very large data bases-Volume. The authors of this article give an example of new model design for qualitative data acquisition, including public administration. The second group insists that the main application of big data in policy making should be a qualitatively new system of information collection based on a set of updating tables, in other words, vitrines collecting information from various spheres of society [7]. An example of the implementation of this approach can serve as an article. Factors influencing big data decision-making quality. Journal of business research, 70, 338-345.". The authors provide not only their own model for organizing big data networks, but also separately assess the potential of using such a model in public administration. It should be noted that such a model can be considered more objective due to a wide range of information sources, but it is extremely difficult to monitor the reliability, timeliness, and formatting of such data [8].

Despite the fact that many works broadcast the idea of the need to use a large set of information, few authors describe its real practical application. Every idea or system must have a commercial or socially positive effect [8], since designing and building even the simplest analytical center will require the use of significant company or institution resources. One example of the full cycle of big data processing can serve the energy industry, which has not only developed a new approach to the evaluation of energy savings, but also building a new management model based on computer forecasting and real-time monitoring [9]. Similar precedents for the introduction of such tools can be observed in other industries, such as urbanism, education, and tourism. In urban planning, a system for gathering already existing statistical information is used, based on which the simulation and forecasting of indicators in the economic field is made - the level of employment, the shortage of residential and commercial space, traffic congestion, and others [10]. In tourism, the forecast is based, among other things, on more real-time information - not only "historical" data, but also the volume of purchased tickets, hotel reservations, actual weather conditions, etc. are considered [11]. The case of using big data in education is different in that when talking about student performance, it is difficult to talk about "historical data" [12]. First, there is often a lack of objective data about students' abilities because evaluation is subjective. Second, multiple years of data may simply not be sufficient to make good predictions about an individual's academic performance. Thirdly, there is a strong personality factor, which can drastically change its vector of educational direction. That is why big data analysis in the field of education of students, for the most part, is based on the data and only indirectly considers the experience of the past.

Consequently, despite its obvious relevance, the comprehensive application of big data for public policy formation is not reflected in existing managerial and scientific practices. That is why the authors of this article set themselves the task of applying information processing tools for forecasting the complex development of regions, to show the effectiveness of using such an approach in state planning and policy making.

One of the components of political planning are state programs [13], intended to improve this or that sphere of public or state life. The way of efficiency planning in Russian practice, as a rule, is reduced to the listing of target indicators. An example is the Unified State

Health Information System which is implemented as part of the national health project in Russia. The project passport specifies the benchmarks to which the result of the project should strive, for example - the reduction of infant mortality to 4.5 cases per 1000 births. However, is it possible to assess the effectiveness of the program not after the fact, but with the help of forecasting tools using IT technology? The authors of [14] believe that the effectiveness of state programs can and should be predictable through modeling tools. Laurence J. O'Toole Jr. и Kenneth J. Meier point to the fact that government programs are usually implemented through complex networks of connected "agents," meaning both different budgetary institutions and private partners. That is why the authors dwell on agent-based modeling in a further study. It is worth noting that the Russian reality is not so characterized using many project implementation agents. Undoubtedly, the tools of private sector involvement are increasingly being used in Russia to solve certain state tasks, but if we turn to the relatively new federal project to create the Unified State Health Information System, we do not see there any decentralized system. Rather the opposite - a clear definition of the supervising agency, its subdivisions, and sever-al small tasks relative to the project itself, the implementation of which is assigned to other agencies [15].

The authors of [16] describe areas of application of simulation modeling in the study of phenomena and processes in management science. Thus, the authors Glenn R. Carroll and Kathleen M. Carley describe examples of the use of simulation modeling for forecasting, making hypotheses and explaining many processes both in the economic life of society and in nature. Although the authors of the paper prove that it is possible to use simulation models (both discrete event and system-dynamic and agent-based) to predict the efficiency of company projects, this paper carries a rather broad generalization of modeling techniques and is difficult to project into the analysis of policy planning efficiency.

The authors of [17] describes in detail the possibility of modeling discrete events on the prediction of further consequences of a particular process. Researchers of [18] present the possibility of implementing a discrete event modelling in the healthcare sphere and suggest that despite several features, modelling in this sphere is not very different from other areas. In the implementation of state programs, there is often a clear set of sequential actions, limited by time, resource and result. Therefore, discrete-event modeling, as part of the implementation of state programs of the Russian Federation, can be used as much as possible for Russia. General conclusions on the available literature on the evaluation of state programs using big data tools, modeling and analysis can be summarized as follows: to date, most of the works on the application of new technologies in public administration are devoted to the development of digitalization and openness of the state apparatus and bureaucratic processes. Quite a few works describe mechanisms for transforming planning processes, policy development, and state programs. One such mechanism could be a modeling tool for forecasting and planning state programs [19] and evaluating their effectiveness [20], but there are virtually no scientific studies that consider the nature of creation and implementation of state programs in Russia in a centralized way [21]. That is why the task of this study, based on a broad and well-developed literature on the use of the modelling mechanism, to use a new system of evaluation and prediction of the effectiveness of existing and potential state projects.

II. Methods

A. Classification of factors

The following groups of indicators were collected and processed during the study for the period of 2010-2020. In total 3064 lines with 419 columns each were used as a data set.

Forecasting the effectiveness of existing and potential State programs could be based on the analysis of big data. The primary research method of this article is modelling. Based on the use of real statistical indicators, a discrete event model will be applied. On the basis of this model, it becomes possible to describe the correlation of Healthcare State Program targets and other factors.

Table I contains information about the 3 most significant indicators in the topic:

TABLE I. GROUPS OF INDICATORS AND THEIR SIGNIFICANCE

General characteristics	Area of the region	0.00099
	Distribution of population by rural and urban areas	0.00475
	Population of the region	0.00474
Regional heads	Previous jobs	5.11216
	Education/academic degree	0.00214
	Length of public service	0.00178
Demographics	Count of births	0.04769
	Natural increase	0.00463
	Number of children whose parents are deprived of parental rights	0.00431
Standard of living	Improvement of living conditions	3.01107
	Real disposable cash income	0.010249795
	Number of doctors	0.0044915676
Legal Aspects	The level of economic crimes	0.02875175
	Number of violations detected	0.003016376
	Number of crimes registered	0.0025703304
Economic Aspects	Financial performance indicator	0.00613
	Gross regional product	0.00171
	The consumer price index	0.00151
Housing	Volume of housing construction	0.00262
	Total floor area of newly commissioned buildings	0.00262
	The total floor area of residential houses commissioned	0.00145
Transportation	Length of paved public roads	0.00311
	Length of public roads	0.00181
	Transported cargo by roads	0.00064
Environmental	Disposed of pollutants	0.40584
	Current (operating) costs of environmental protection	0.00328
	Electricity consumption	0.00230
Business	Turnover of medium-sized organizations	0.09873
	Turnover of financial and credit organizations	0.08370
	Turnover of small and micro enterprises	0.02136
Labor market	The number of researchers with a doctoral degree	0.01554
	Labor intensity coefficient	0.00239
	Number of employed at the age of 15-72 years	0.00222
Environmental situation	Number of emergencies	0.00556
	Number of people killed in emergencies	0.00191
	Air pollutant emissions from stationary sources	0.00180
Technological indicators	Share of investments in reconstruction and modernization in the total volume of investments in fixed assets	0.00045
	Advanced production technologies in use	0.00080
	Internal current expenditures on research and development by cost type	0.00061

B. Simulation modelling

Infant mortality was selected as an endpoint that characterizes a quality of a regional healthcare system. We performed a Gradient Boost search to find the best performing model with a 5-fold cross-validation based on the training (80%) and testing (20%) datasets. We performed a correlation analysis using F-Measure to find the predictors with the highest predictive power.

III. RESULTS

In the course of the study, a regression model of gradient boosting was developed for the "Infant mortality" target. The median was used for the strategy of filling in the gaps.

Regularization was also applied, the best parameters are listed below (xgb_reg = xgb.XGBRegressor (max_depth = 3, n_estimators = 100, n_jobs = 2, objective ='reg:squarederror', booster='gbtree', random_state = 42, learning_rate = 0.05). Next, the significance of the predictors was calculated by the SHAP values. Fig. 1 shows the SHAP values of the predictors.

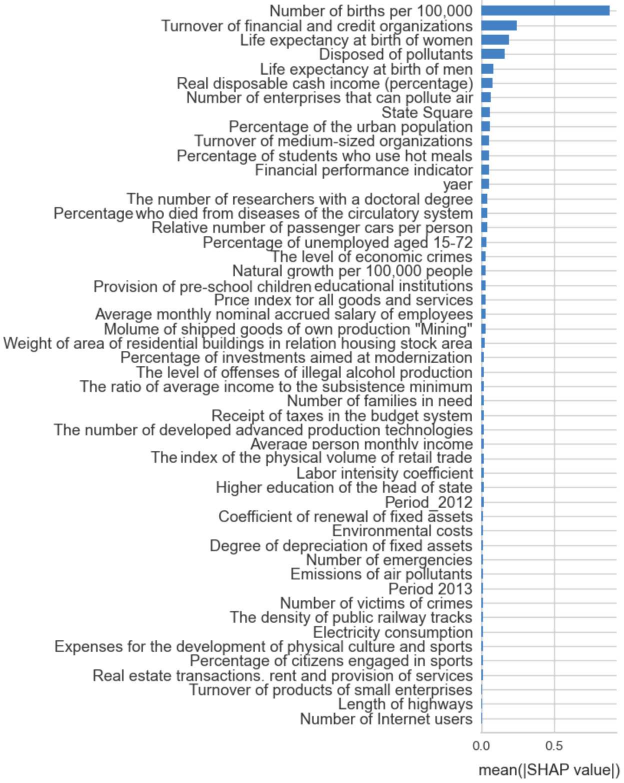

Fig. 1. SHAP values of the predictors

The results show both quite expected connections and not so obvious ones. First, a group of statistically related indicators is observed. Thus, the greatest im-portance of the indicator "number of births (without stillbirths)" (F score 217) is obvious, because an increase in the number of births is bound to lead to a certain increase in infant mortality.

The same relationship is observed with the indicator "natural increase for the year" (F score 39), because natural increase (NI) is determined by the formula that considers the number of births (as mentioned above):

$$NI = B(t) – M(t), \qquad (1)$$

where B(t) — number of births in a year t (i.e. the index with F score 120), M(t) — number of deaths in a year t.

The same relationship is observed with the indicator "life expectancy at birth in men" (F score 86), because this indicator is determined by a formula that considers infant mortality:

$$e_0 = \sum_{k=0}^{\infty}\prod_{i=0}^{k} p_i \qquad (2)$$

where pi — is the probability of living to the end of the year for a man of age i at the beginning of the year, determined by the formula: pi = 1 – qi , where qi — is the proportion of deaths before the end of the year of persons of age i at the beginning of the year according to the mortality table for that year, published after the end of the year.

Fig. 2 shows the results of a random forest regression model features importance for the infant mortality.

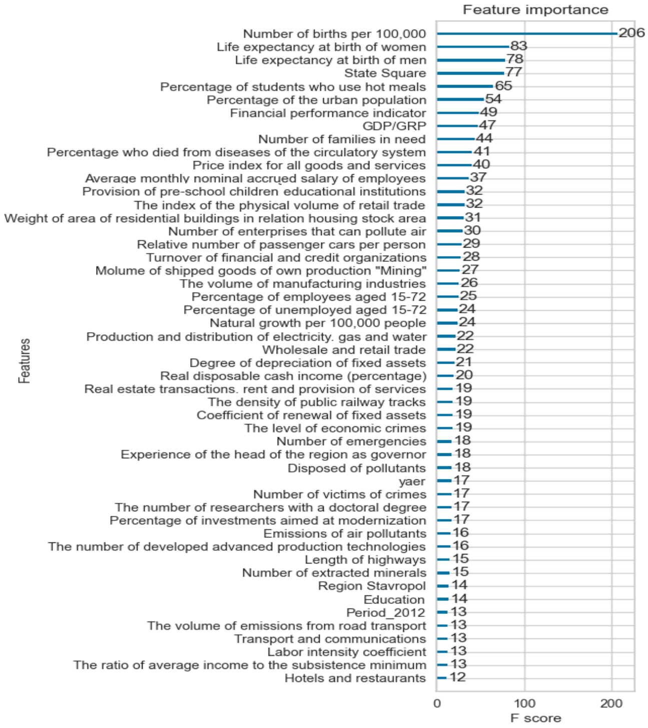

Fig. 2. F-measures features importance for infant mortality

At the same time, the indicator "life expectancy at birth for women" is expected to be somewhat lower (F score 79), because infant mortality for girls is significantly lower than for boys, both in Russia and in the world.

Second, there are indicators that generally reflect the level of social welfare. The results show which of the indicators of this kind turn out to be the most significant, and therein lies the value of the extracted social knowledge.

Obviously, the "number of families registered as needy" (F score 20), the "number of emergencies" (F score 18) in some cases directly affect the economic or physical ability of the population to prevent infant mortality.

Fig. 3 Shows a decision tree for an infant mortality

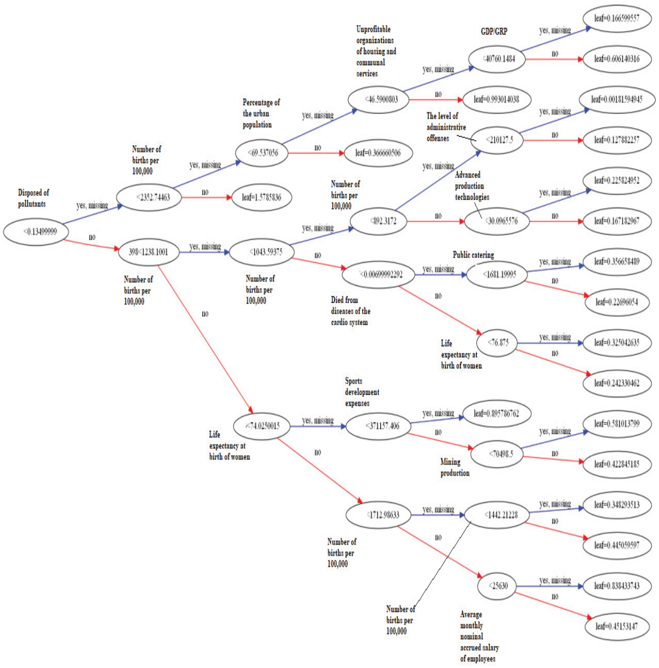

Fig. 3. Decision tree regression model

Third, several indicators reflect three trends, namely, the nexus of infant mortality:

(a) With the area of the territory (F score 56);

(b) With the share of urban population (F score 58) and closely related indicators "proportion of students with hot meals" (F score 62), "provision of children with places in pre-school institutions" (F score 29) (F score 29), "real disposable income" (F score 27), "wholesale and retail trade (F score 22), "number of own cars" (F score 21), "production and distribution of electricity, gas and water" (F score 20), "number of employees" (F score 19);

(c) With economic activity - "financial activity" (F score 48), "GDP/GRP" (F score 27), "investment for reconstruction and modernization" (F score 25), "wages of employees" (F score 22), "fixed assets renewal rate" (F score (F score 22), "number of developed advanced production technologies" (F score 20) (F score 20), "fishing, fish farming" (F score 19), etc.

The resulting model performance was R_squared = 0.9874880265442392 Fig. 4.

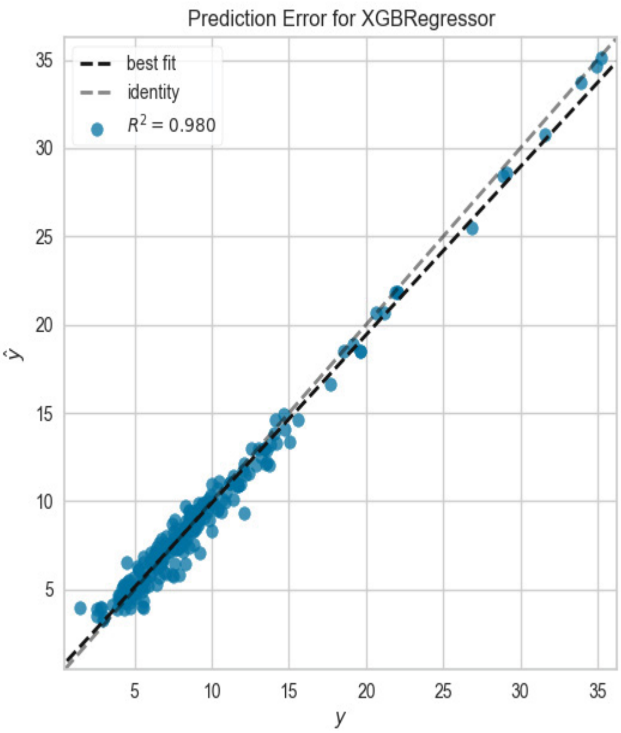

Fig. 4. Prediction error plot for infant mortality

Root mean square error = 0.47. The graph shows the distribution of the squared error, indicating the high accuracy of the model. Prediction error plot is shown in the Fig. 5.

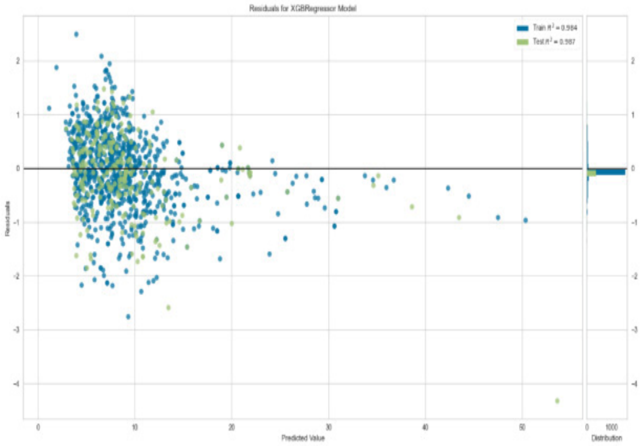

Fig. 5. Residual plot for the prediction

IV. Conclusions

During the research work, it was possible to identify factors affecting the indicators of the national health project. It is worth noting that the indicators of national projects are influenced by the level of education of the head of the region and work experience. The combination of a large territory, a high proportion of the urban population and economic activity are three important factors that require the modernization of the system of public administration. However, in Russian conditions, when public administration is highly unified, such a transformation practically does not take place. These negative trends are manifested in the economy ("consumer price index in relation to the previous year" - F score 38), the social sphere ("the share of deaths from circulatory system diseases" - F score 32) and the rule of law ("the number of economic crimes" - F score 20). It is indicative that there is a correlation with another managerial indicator, which in a democracy usually has a negative connotation - "length of service as governor" (F score 24). The number of economic crimes - F score 20, also positively influence the growth of infant mortality.

The contribution to the top indicators of national projects by category is shown in Table II.

TABLE II. Total count and value of the predictors and it Weight

Features category	Features Count	Features importance
Project features	4	14
Technological	3	13
Transport	5	12
Housing	4	11
Environmental situation	5	13
Economic Aspects	9	9
Region characteristics	4	8
Labor market	8	7
Legal Aspects	27	6
Demographics	20	5
Business	20	4
Environment parameter	5	3
Standard of living	26	2
Regional heads	10	1

The table shows that the most affected indicators of the health program are groups such as: regional heads (previous job, previous professional experience, specialization, level of education, academic degree, etc.), standard of living (housing conditions, income per person, number of doctors in the region, etc.), and environmental indicators (pollution, funding level of government programs, electricity consumption, etc.). Other categories had significantly less influence on the dynamics of the target indicators. However, it is possible to highlight: the business category, the demographic category, as well as the legal aspect, and the labor market. These groups follow the above mentioned "leaders" of influence.

It is revealing that there is a correlation with two other economic and managerial indicators: infant mortality correlates positively with "years in office as governor" (F score 24) and negatively with "GDP/GDP" (F score 27). In other words, the shorter a governor's term in office, the smaller the territory under control, percentage of the urban population, and economic activity (including retail and wholesale trade, transportation of cargo, vehicle use, etc.), but the higher the gross regional product, the lower the infant mortality rate.

Thus, turnover of power, standard conditions of governance (territory, population, and economy), and a large gross regional product are the three main factors that reduce infant mortality.

In this model, the heterogeneity of big data provides complex analysis. The application of this model to forecasting and analysis in the development and implementation of government programs makes it possible to supplement program-objective methods with more fine and accurate regulatory tools to achieve greater efficiency of government programs.

In future research application of the modeling in other spheres can provide a comprehensive overview of the mechanisms of realization any state programs. Depending on specific target indicators one can carry out an analogous research and find out which parameters are responsible for the trend of the target indicators. Moreover, it is possible to transpose the research focus to the human level and determine how one or another personal characteristic (for example – the governor of the region) affects the socioeconomic progress of the territory. The application of the methodology described in this article has a huge potential for future research.

Such regulatory tools can be varied and based on predictive model data. For example, laws on the term of office of senior regional officials or laws on the re-election of these officials may be changed. There could also be programs to provide the required infrastructure to the suburbs and villages, to motivate the urban population to live in the countryside, to take care of their environment. Programs for GDP growth are being adopted as they are but there is a possibility that the theoretical approaches to achieving economic growth need to be revised. The use of big "legal" data (for example, on offenses and crimes) and big data in the political, administrative, budgetary, social, economic, and environmental fields in the model allows revealing implicit complex relationships between processes and indicators. The knowledge of these links has great potential for the introduction of regulatory sandboxes to have a targeted impact on individual processes and indicators to achieve the goals of government programs. Thus, this model can provide new knowledge to optimize legal regulation and improve government programs.

Acknowledgment

The reported study was funded by RFBR, project number 20-011-00837.

References

[1] J. Höchtl, P. Parycek, and R. Schöllhammer, “Big data in the policy cycle: Policy decision making in the digital era,” https://doi.org/10.1080/10919392.2015.1125187, vol. 26, no. 1–2, pp. 147–169, Apr. 2016, doi: 10.1080/10919392.2015.1125187.

[2] Lee-Geiller, S., & Lee, T. D. (2019). Using government websites to enhance democratic E-governance: A conceptual model for evaluation. Government Information Quarterly, 36(2), 208-225.

[3] Millard, J. (2017). European Strategies for e-Governance to 2020 and Beyond. In Government 3.0–Next Generation Government Technology Infrastructure and Services (pp. 1-25). Springer, Cham.

[4] Ju, J., Liu, L., & Feng, Y. (2019). Public and private value in citizen participation in E-governance: Evidence from a government-sponsored green commuting platform. Government Information Quarterly, 36(4), 101400.

[5] Engin, Z., & Treleaven, P. (2019). Algorithmic government: Automating public services and supporting civil servants in using data science technologies. The Computer Journal, 62(3), 448-460.

[6] Deshpande, A., Guestrin, C., Madden, S. R., Hellerstein, J. M., & Hong, W. (2004, August). Model-driven data acquisition in sensor networks. In Proceedings of the Thirtieth international conference on Very large data bases-Volume 30 (pp. 588-599).

[7] Janssen, M., van der Voort, H., & Wahyudi, A. (2017). Factors influencing big data decision-making quality. Journal of business research, 70, 338-345.

[8] Mikalef, P., Boura, M., Lekakos, G., & Krogstie, J. (2019). Big data analytics and firm performance: Findings from a mixed-method approach. Journal of Business Research, 98, 261-276.

[9] Zhou, K., Fu, C., & Yang, S. (2016). Big data driven smart energy management: From big data to big insights. Renewable and Sustainable Energy Reviews, 56, 215-225.

[10] Silva, B. N., Khan, M., & Han, K. (2018). Towards sustainable smart cities: A review of trends, architectures, components, and open challenges in smart cities. Sustainable Cities and Society, 38, 697-713.

[11] Ivars-Baidal, J. A., Celdrán-Bernabeu, M. A., Mazón, J. N., & Perles-Ivars, Á. F. (2019). Smart destinations and the evolution of ICTs: a new scenario for destination management? Current Issues in Tourism, 22(13), 1581-1600.

[12] Williamson, B. (2016). Digital education governance: data visualization, predictive analytics, and ‘real-time’ policy instruments. Journal of Education Policy, 31(2), 123-141.

[13] Denhardt, R. B. (1985). Strategic planning in state and local government. State & Local Government Review, 174-179.

[14] O'Toole Jr, L. J., & Meier, K. J. (1999). Modeling the impact of public management: Implications of structural context. Journal of public administration research and theory, 9(4), 505-526.

[15] O. Metsker, G. Kopanitsa, and E. Bolgova, “Prediction of childbirth mortality using machine learning,” 2020. doi: 10.3233/SHTI200623.

[16] Harrison, J. R., Lin, Z., Carroll, G. R., & Carley, K. M. (2007). Simulation modeling in organizational and management research. Academy of management review, 32(4), 1229-1245.

[17] G. S. Fishman, “Simulation in Perspective,” Discrete-Event Simulation, pp. 1–35, 2001, doi: 10.1007/978-1-4757-3552-9_1.

[18] M. M. Günal and M. Pidd, “Discrete event simulation for performance modelling in health care: a review of the literature,” Journal of Simulation 2010 4:1, vol. 4, no. 1, pp. 42–51, Mar. 2010, doi: 10.1057/JOS.2009.25.

[19] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260..

[20] Debnath, R., & Bardhan, R. (2020). India nudges to contain COVID-19 pandemic: A reactive public policy analysis using machine-learning based topic modelling. *PLoS One*, *15*(9), e0238972.

[21] Gareev, M. (2020). Use of Machine Learning Methods to Forecast Investment in Russia. Russian Journal of Money and Finance, 79(1), 35-56.