DOI: 10.15276/aait.01.2021.7
UDC 004.02+004.62+004.9

FRAMEWORK FOR SYSTEMATIZATION OF DATA SCIENCE METHODS

Vira V. Liubchenko[1)]
ORCID: http://orcid.org/0000-0002-4611-7832; lvv@opu.ua
Nataliia O. Komleva[1)]
ORCID: http://orcid.org/0000-0001-9627-8530; komleva@opu.ua
Svitlana L. Zinovatna[1)]
ORCID: http://orcid.org/0000-0002-9190-6486; zinovatnaya.svetlana@opu.ua
Katherine O. Pysarenko[1)]
ORCID: https://orcid.org/0000-0001-9573-9315, horodnychaya@opu.ua
[1)] Odesa National Polytechnic University, 1, Shevchenko Ave, Odesa, 65044, Ukraine

ABSTRACT

The rapid development of data science has led to the accumulation of many models, methods, and techniques that had been successfully applied. As the analysis of publications has shown, the systematization of data science methods and techniques is an urgent task. However, in most cases, the results are relevant to applications in a particular problem domain. The paper develops the framework for the systematization of data science methods, neither domain-oriented nor task-oriented. The metamodel-method-technique hierarchy organizes the relationships between existing methods and techniques and reduces the complexity of their understanding. The first level of the hierarchy consists of metamodels of data preprocessing, data modeling, and data visualization. The second level comprises methods corresponded to metamodels. The third level collects the main techniques grouped according to methods. The authors describe the guiding principles of the framework use. It provides a possibility to define the typical process of problem-solving with data science methods. A case study is used to verify the framework's appropriateness. Four cases of applying data science methods to solve practical problems described in publications are examined. It is shown that the described solutions are entirely agreed with the proposed framework. The recommended directions for applying the framework are defined. The constraint of the framework applying is structured or semi-structured data that should be analyzed. Finally, the ways of further research are given.

Keywords: Data science; framework; data preprocessing; data modeling; data visualization; case study

For citation: **Liubchenko V., Komleva N., Zinovatna S., Pysarenko K. Framework for Systematization of Data Science Methods.** ***Applied Aspects of Information Technology.*** **2021; Vol.4 No.1: 80–90. DOI: 10.15276/aait.01.2021.7**

INTRODUCTION. FORMULATION OF THE PROBLEM

Data science is a rapidly evolving field, which combines problem domain knowledge, programming skills, and knowledge in mathematics and statistics to extract information and insights from data.

Data science is a discipline, which considers data exploration and preparation, data representation and transformation, computing with data, data modeling, data visualization and presentation. The results of data science applications are valuable for business. They can guide how to increase revenue, reduce operating costs, find anomalies, determine strategy, etc.

According to [1], the development of a business software system includes several iterative stages: business understanding, data understanding, data preparation, data modeling, performance evaluation, and system deployment. That is the work with data associates with data preparation, modeling, and presentation to users.

The main stages of the fundamental methodology of data science were detailed described in [2]. Data preparation is realized after the data requirements formulation, which depends on the selected analytical approaches, data collection, when "data scientists identify and gather the available data resources – structured, unstructured and semi-structured – relevant to the problem domain". The development of new methods and techniques, implemented by various software tools, provides researchers and practitioners with a wide field to process and analyze available data. However, selecting the most suitable methods and tools remains a non-trivial task, the solution of which is not formalized and is based on the researcher's experience and intuition. Therefore, the systematization of data science methods in terms of their suitability for particular problems solving is actual.

The current paper aims to provide a reasonable systematization of data science methods in the form of a framework, which simplifies the choice of relevant techniques for solving specific problems.

For achieving this aim, the paper solves the following tasks:

1) analysis of typical techniques of data science and their generalization in the form of a metamodel–method–technique structure;

2) description of the guiding principles for framework use;

3) validation of the fitting the framework-based recommendation to the choices of data science techniques in real cases;

4) discussion of opportunities, limitations, and open challenges for the proposed framework.

1. RELATED WORK

Interest in data science has been consistently high in recent years (Fig. 1).

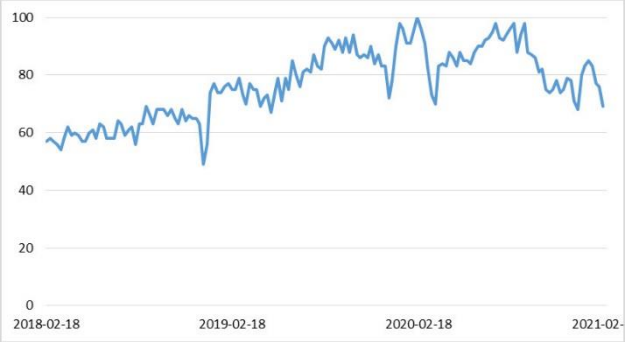

Fig. 1. **Hit count on term "Data Science" in the last three years (data collected on 18th February 2021)**
Source: Google Trends (www.google.com/trends)

Data science is an open science system that involves many stakeholders from various institutions around the world. Researchers and practitioners have developed elements of this system following their own goals and interests, which has led to the accumulation of many models, methods, algorithms, technologies. The data science ecosystem works stably, but due to a large amount of contained information, systematizing this information becomes valuable.

We note that many publications provide an overview of various aspects of data science. There are reviews on specific characteristics, such as the history of development and terminology [3] or the factors influential for data science projects' success [4].

However, most reviews address specific groups of methods or issues in applying data science to specific subject areas.

The review [5] presented a data science conceptual map that visualizes the relationship between data science concepts. However, due to the need to consider a wide range of issues in the picture, only model-based methods of working with data are presented, but their detail is not provided.

The review [6] focused on data science's fundamental concepts for solving cultural artifacts comparison problems. The review is descriptive and does not provide a detailed description of the recommended techniques.

Review of data science opportunities for managerial research [7] focused is on five areas: collection, storage, processing, analysis, and reporting. However, as in the previous case, a high-level description is provided without detail. Due to the peculiarities of the problem domain, attention is focused only on regression methods.

In [8], the advantages and disadvantages of different modeling methods were considered. However, the analysis is limited to using only two modeling options: data modeling and simulation. Also, these options are considered only for big data, which reduces the scope of the application.

Paper [9] provided a thorough description of big data applications in e-commerce and identified areas for further research, identified the types of data used in e-commerce applications and the values that businesses can derive from big data analysis. However, the methods of data analysis and their systematization remained out of the authors' attention.

In [10], big data analysis methods used in supply chain management are considered. This work is precious because it developed a framework for methods classification based on the analysis of publications. The limitation of the framework use is caused by basing on the tasks of supply chain management.

The authors [11] provided a systematic overview of data science techniques used to process game learning analytical data. They grouped all techniques into three categories: supervised learning models, unsupervised learning models, and visualization. However, the first and second categories combine techniques that solve different problems. The relationships between a category of techniques and solved problems are not reflected in work.

Data science is actively used to analyze medical data, but relevant publications usually focus on specific problems to be solved. For example, [12] considered classification techniques in the prediction of chronic diseases, [13] reviewed the data science techniques used to monitor health and support to cancer patients managed at home, and [14] presented the application of data mining techniques for the mental illness diagnosis.

In [15], the authors reviewed approaches to diagnosing the bridges' structural health based on the collected data. Analysis techniques were selected according to the problem's context. They covered the compressive sampling-based data-acquisition algorithm, the anomaly data diagnosis approach using a deep learning algorithm, crack identification approaches using computer vision techniques, and the condition assessment approach for bridges using machine-learning algorithms.

The work [16] attracted our attention because its authors, based on the analysis of publications and their own research, had shown how the combination of machine learning techniques and data visualization boosts sense-making and analytical reasoning through mutual enrichment of opportunities. However, the work is performed in the human-machine interaction field and categorized types of algorithms and degree of interactivity only. Therefore, the proposed systematization has limited application.

Data visualization is a crucial stage in data-driven research. The customer of any data science product is business. Data visualization is the key success factor for communication with business [1]. For leading to valuable results, data science products should be understandable for not-technically skilled personnel.

The work [17] also noted: "data visualizations must be intuitive for technical or non-technical users to assume it and draw significant insight from it." Visualization is a powerful tool "to figure out valuable patterns, learn from historical data, and find the best way forward for success."

The selection of data visualization tools should be based on knowledge about volume, variety, velocity, and value of analyzed data [18].

The visualization tools should meet requirements categorized as dimensionality reduction, data reduction, scalability and readability, interactivity, fast retrieval of results, and user assistance [19].

The publications also paid attention to the fact that data cleansing and preparation are essential for any data science application. The paper [20] described three categories of data anomalies: syntactic anomalies, semantic anomalies, and coverage anomalies. The work [21] contained a list of data preparation techniques but provided description has not confirmed its completeness. The authors of [22, 23] described general problems of data cleansing and transformation, but they emphasized the task of testing the ETL (Extract, Transform, and Load) process in data warehouses. The review [24] described data cleaning methods, but it is strongly oriented on web information systems.

Thus, the publications described many different methods and algorithms as well as cases of their application. However, published reviews are sometimes too generic and, in most cases, focused on solving problems of a specific domain. Because of this, information about data science methods gathered during years is unstructured and difficult to understand. We apply hierarchical decomposition as a means of overcoming the system complexity [25].

2. FRAMEWORK DESIGN

Before we start designing a framework, we have had to formulate requirements for it. First, the framework should reflect and summarize the accumulated practical experience. Second, it should also be intuitive to support the decision on the method selection in particular conditions. Finally, the framework should be resilient to further expansion by new knowledge addition.

We believe that it is appropriate to take a three-level hierarchy: the highest level represents metamodels corresponded to specific purposes, the middle level lists methods that solve a specific problem, and the lowest level enumerates specific techniques for solving the problem.

The analysis of the published cases and the best practices allows offering to distinguish three types of the purposes of data science applications.

Data preprocessing is the process of preparing raw data for further work. We can describe the necessary transformations by the metamodel

$$data = f_p(raw\_data),$$

where *raw_data* and *data* represent raw and prepared data, respectively, f_p is a conversion function.

Data modeling is the process of producing new knowledge based on data. The corresponding metamodel is

$$insight = f_m(data),$$

where f_m is a function that describes model transformations, *insight* represents generated description of new knowledge.

Data visualization is the process of providing a visual representation of data to facilitate their understanding. The corresponding metamodel is

$$plot = f_d(data),$$

where f_d is a function that specifies the type of visual rendering, *plot* is a presentation of data as a one-, two- or three-dimensional chart.

Three types of methods can be distinguished among a set of data preprocessing methods.

The methods of the first type should clean-up the raw data. The real-world data collected for analysis is dirty; they contain low-quality records. "Garbage in, garbage out" is the fundamental principle of data analysis. Therefore, raw data should be processed to ensure that the garbage never gets in the first place.

Therefore, the raw data are the inputs of cleansing methods, and the parameters are the error types and the method of removing errors. The common types of errors are missing values, duplicated records, impossible values, outliers, etc. Records with

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

errors could be dropped or corrected according to a specified rule.

The methods of the second type transform data, so it takes a suitable form for data modeling. Normalization and scaling usually are used to ensure feature compatibility. Also, the transformation could shape data to another measurement scale, aggregate features, etc.

The methods of the third type relate to feature engineering and improve the performance of the further analysis. They concern the dimensionality reducing, additional data providing, binning, etc.

Both raw data and cleaned data could be input for preprocessing methods of second and third types; the parameters are the types of required transformations. The preprocessing method usually consists of five steps.

1. Select the input data.
2. Clear-up the input data.
3. Determine the type of processing required.
4. Perform the necessary processing.
5. Get descriptive characteristics of the data.

The set of data modeling methods could be differed according to the time horizon for insight. In generative modeling, one proposes a stochastic model that could have generated the data. The last count of available data limits the time horizon. In predictive modeling, one constructs methods that predict well over some given data universe. The time horizon for prediction is defined by the analyst and restricted by the predictive possibility of a model.

The data after preprocessing is the input for data modeling methods; the parameters are variables that should be included in the model.

Data modeling methods usually consist of four steps.

1. Selection of variable to enter in the model.
2. Selection of appropriate modeling technique.
3. Execution of the model.
4. Diagnosis and model comparison.

Data visualization can help understand data and the interactions between variables, support formulation hypothesis for data analysis, and provide results demonstration. Data presentation methods are designed for experienced analysts and can support exploratory data analysis and visualize the results obtained in data modeling. Interactive visualization methods support exploratory data analysis for less technically skilled, more application-oriented personnel.

The visualization method's input is either raw data or data after preprocessing or the applied models' results; the parameter is the type of drawn chart. Unlike other methods, parameter specification is optional. Chart types can be determined automatically based on the measurement scale type.

The visualization method consists of three steps.

1. Read the input data.
2. Identify the charts that are appropriate to create.
3. Create the necessary charts and conclude.

Fig. 2 shows the "metamodel – method" levels of the proposed framework.

Now move on to the next level of the hierarchy – the level of techniques.

Data cleansing techniques depend on the type of error they need to remove and the available data volume. The common types of errors are missing values, impossible values, outliers; respective preprocessing techniques are observation removing or values replacement. If the observation contains errors, it can be removed. However, in this case, the observation information would be lost, and this technique can significantly reduce the sample size. In the case of missing or impossible values, they could be replaced by "neutral" values, such as the mean or mode. In the case of data type mismatch, lack of decimal point, extra zeros, etc., type conversion or parsing with regular expressions should be used.

Data transformation techniques change the data format or value to a particular format and presentation to meet the analytical model's assumptions. The standardization technique scales data to establish the mean and the standard deviation at 0 and 1, respectively, by transformation

$$x_s = \frac{x - \overline{x}}{\sigma},$$

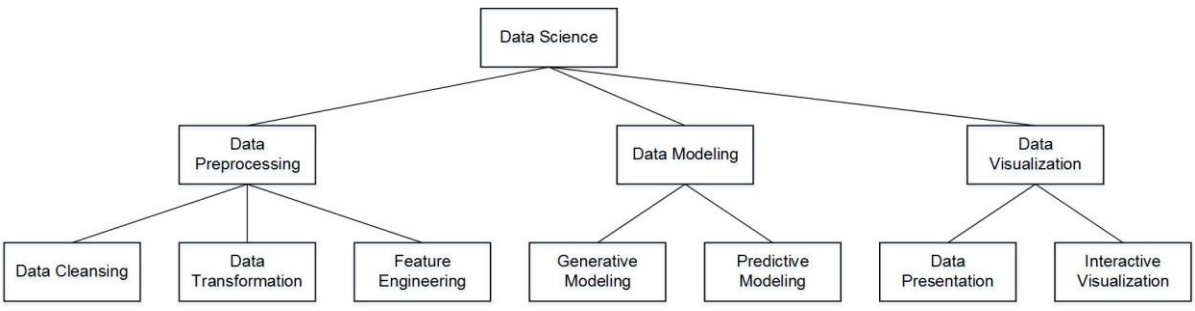

Fig. 2. **Levels "metamodel – method" of the framework**
***Source*: Compiled by the author**

where x is the attribute's initial value, $\bar{x}$ and σ are the mean and standard deviation of the attribute's initial values. As a result of standardization, negative values can be obtained. Sometimes negative values do not fit the model's assumption. The scaling technique scales data to any given range. For example, scaling to the range [0…1] is realized as follows

$$x_n = \frac{x - \min(x)}{\max(x) - \min(x)},$$

where x_n is a scaled value, min(x) and max(x) are minimal and maximal attribute's initial values.

Analysis methods are usually sensitive to the data scale; the type of measure scale affects the analysis methods and data understanding. Data encoding techniques allow converting categorical data to numeric. Sometimes there is a need to perform mathematical transformations of data, such as log-transformation.

The data may be gathered from multiple data sources. In such case, they should be combined these into a summary for data analysis. Data aggregation techniques gather data and present it in a summarized format.

Feature engineering changes the structure of the data. Sometimes there are too many attributes, so one needs to reduce the number because they do not add new information to the model. In this case, use the technique of dimensionality reduction. Sometimes, there is a need to group similar data; for this purpose, bucketing technique places similar values into buckets. Also, sometimes there is a need to create additional data based on the input sample, which slightly differs from it; the augmentation technique is used for this purpose.

The levels "method – technique" for preprocessing methods are shown in Fig. 3.

Modeling techniques are often divided into statistical techniques, tree techniques, and machine learning techniques. However, such classification does not meet our requirements due to insufficient accuracy. Also, it does not provide unambiguity for selection, as several techniques can be used to model data for the same purpose in many cases. So we link techniques to particular tasks and leave the implementation means beyond the hierarchy (Fig. 4).

Data exploring techniques are designed to gain an understanding of data and the interactions between variables. Usually, data exploring techniques concern the calculation of descriptive statistics and are followed by visualization techniques.

Regression analysis techniques are designed to identify the shape of a relationship between numerical predictor variables and the resulting variable. In data science, the purpose of applying this technique in most cases is to predict the numerical results. Therefore, the implementation means can be both statistical methods and neural networks. However, if the task is to describe a particular phenomenon, only statistical methods are used as a mean.

Classification techniques are designed to predict the category for a new observation. Classification techniques are techniques of supervised learning and use labeled training data.

Clustering techniques are designed to organize similar observations into clusters and are techniques of unsupervised learning. They can also identify typical representatives for each identified cluster.

The levels "method – technique" for modeling methods are shown in Fig. 4.

Data presentation techniques depend on the nature of data and the purpose of visual presentation. Three types of charts are most often used. Bar charts are used to show how some quantity varies among some discrete set of items or for plotting histograms of bucketed numeric values to explore how the values are distributed visually. Line charts are used to show trends. Scattering diagrams are used to visualize the relationship between two paired sets of data. Box plots are used to display numerical data groups and are convenient for datasets comparison and extreme values visualizing. Heat maps are used for multidimensional data analysis and correlation detection.

Interactive visualization is associated with the construction of dashboards, which simplify the interaction of a less technically skilled user with data sets. Typically, dashboards are designed using data presentation techniques.

The levels "method – technique" for visualization methods are shown in Fig. 5.

The proposed three-level hierarchy is static. It formalizes the existing links between data science methods and techniques but does not provide recommendations for its application. Now we describe how to apply the framework.

3. GUIDING PRINCIPLES FOR THE FRAMEWORK USE

The problem T solution, regardless of its complexity and the problem domain, is an achievement of particular aim A, which usually consists of a set of N tasks:

$$A = \bigcup_{i=1}^{N} a_i,$$

where each a_i is a separate research task.

The process of problem resolving contains the sequence of phases p_j, $p_j \in P$, each of them responds for solving a subset of tasks A_j:

$$\exists a_i, mmt_k \left| a_i \in A_j \wedge solvable(a_i, mmt_k) = true \wedge mmt_k \in MMT_j \right.$$

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

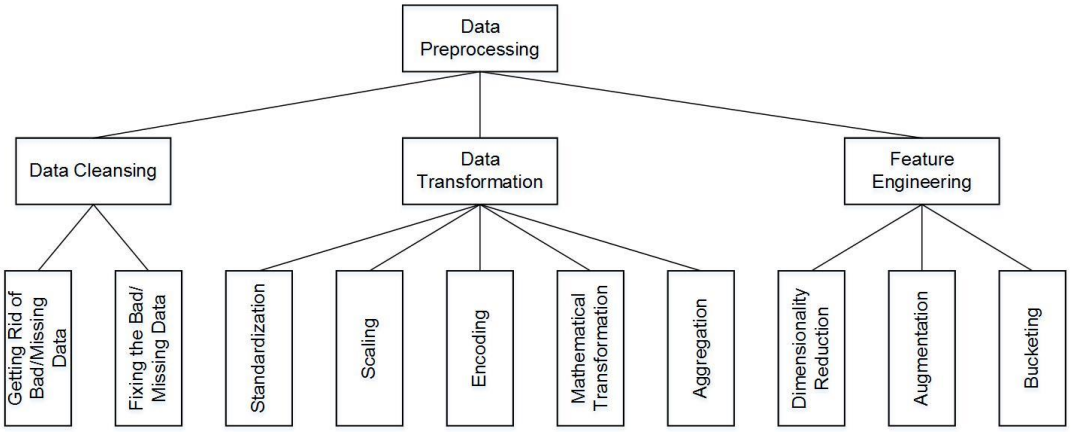

Fig. 3. **Levels "method – technique" for preprocessing methods of the framework**
***Source*: Compiled by the author**

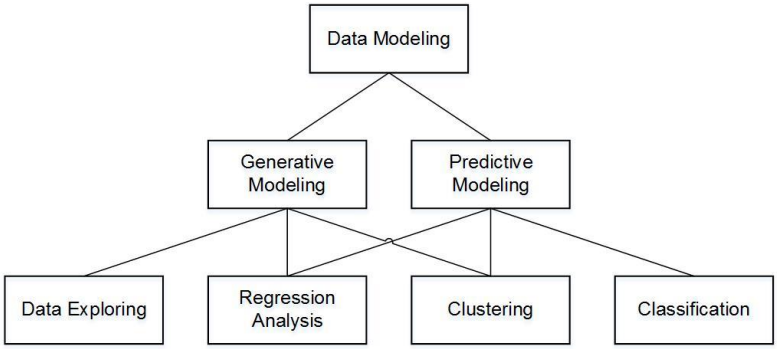

Fig. 4. **Levels "method – technique" for modeling methods of the framework**
***Source*: Compiled by the author**

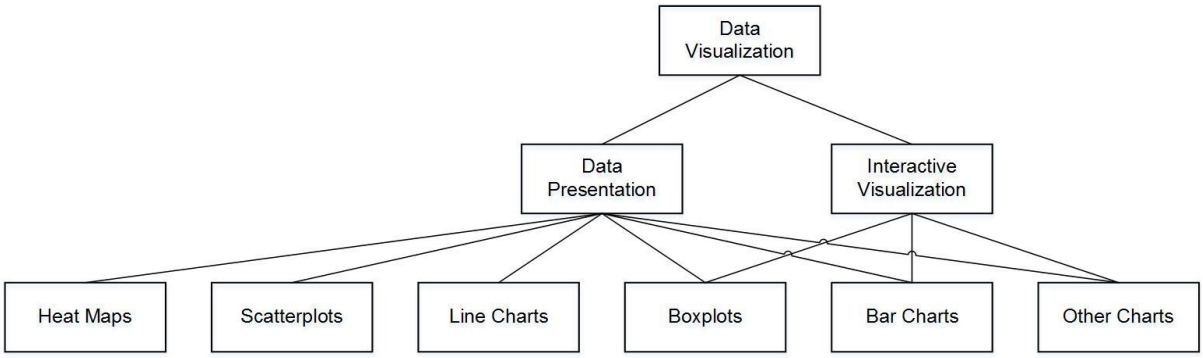

Fig. 5. **Levels "method – technique" for visualization methods of the framework**
***Source*: Compiled by the author**

where *solvable*(a_i,mmt_k) is a predicate, which takes true if a triple mmt_k should be used to succeed in task a_i, MMT_j is a set of triples "metamodel – method – technique" for the phase p_j.

For purposes of this paper, the following terms have the following meanings.

Two phases p_1 and p_2 are *aim-independent* if $A_1 \cap A_2 = \varnothing$.

Two phases p_1 and p_2 are *time-independent* if $t_1 \cap t_2 = \varnothing, t_1 = \left\langle t_1^{\min}, t_1^{\max} \right\rangle, t_2 = \left\langle t_2^{\min}, t_2^{\max} \right\rangle$.

The problem-solving process is *manageable* if the number of unsolved tasks decreases over time: $|A_1| > |A_2|$ if $t_1^{\min} \le t_2^{\min}$ and $t_1^{\max} \le t_2^{\max}$, i.e., tasks A_1 should be solved before the solving of tasks A_2 starts.

Two phases p_1 and p_2 are *structure-independent* if $MMT_1 \cap MMT_2 = \varnothing$.

The algorithm of the framework using for problem T solving consists of five steps.

1. Decompose the aim A on the tasks depending on the semantics of problem T.

2. Group tasks into phases P depending on the technological aspects of problem T solving. Establish relationships between phases and define their time order.

3. Define the set MMT_i for each phase p_i following the framework (Fig. 2–5).

4. Execute all phases following the defined techniques.

5. If the set of unsolved tasks is not empty, then return to step 1; otherwise – end.

As a result of the framework using, data scientist gets the work plan as a set of quadruples $\left\langle p_i, A_i, MMT_i, \left\langle t_i^{\min}, t_i^{\max} \right\rangle \right\rangle$. Note that a complete plan does not have to be obtained before starting work. When the steps are structure-independent, the final definition of the methods and techniques in the triplets MMT_i could be postponed.

4. CASE STUDY

To assess the proposed framework's compliance with the experience, we have studied published cases.

As the first case, consider the actual problem of predicting the situation with COVID-19. Publications that differ by country, forecast, and model appear daily. As a case for analysis, we have taken the solution described in [26]. The prediction of disease incidence was performed based on data on the disease in previous days and the number of queries for specific keywords from Google Trends.

The study aimed to predict the incidence of COVID-19 in a particular geographical region without using personalized data.

Following the guiding principles for framework use, the aim can be presented as

$$A=<a_1, a_2, a_3, a_4, a_5>,$$

where a_1 is a task to obtain data on daily new cases of coronavirus; a_2 is a task to obtain data on search interest in particular terms from Google Trends; a_3 is a task of data integration; a_4 is a task to build a predictive model with a given accuracy; a_5 is a task to predict the incidence of COVID-19 in the country.

Tasks grouping places them in two phases: p_1 – data preprocessing, p_2 – predictive modeling. The phases are time-, aim- and structure-independent. A network diagram of the process is shown in Fig. 6.

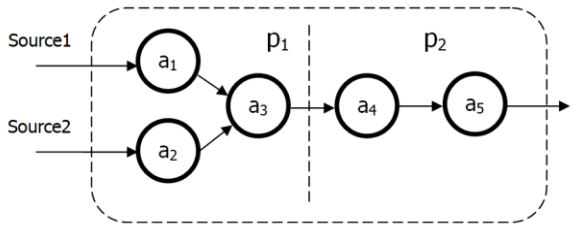

Fig. 6. **Dividing the problem solving into phases**
***Source*: Compiled by the author**

At the first phase, data preprocessing is performed $<p_1, A_1, MMT_1, <t_{1.1}, t_{1.2}>>$, where $A_1=<a_1, a_2, a_3>$.

Solving tasks a_1 and a_2 could be performed in parallel. Both tasks aim to ensure the relevance of data obtained from respective information sources. Tasks $\{a_1, a_2\}$ and a_3 are sequential in time. The authors of [26] believed, datasets obtained from Worldometer are reliable and do not require cleansing. So $MMT_1 = <mmt_1>$, mmt_1 is data aggregation, which should combine and synchronize data from two sources.

At the second phase, data predictive modeling is performed $<p_2, A_2, MMT_2, < t_{2.1}, t_{2.2}>>$, $A_2=<a_4, a_5>$, $MMT_2 = <mmt_2>$, mmt_2 is regression analysis, $t_{1.1} < t_{2.1}$ and $t_{1.2} < t_{2.2}$. Tasks a_4 and a_5 are sequential in time. Since the aim is an accurate prediction of the incidence, not the analysis of the model of particular factors' impact, the regression analysis was implemented with a recurrent neural network.

The sequence of applied methods for predicting the incidence of COVID-19 is shown in the generalized form in Fig. 7. Note the compliance of the case with the framework application result.

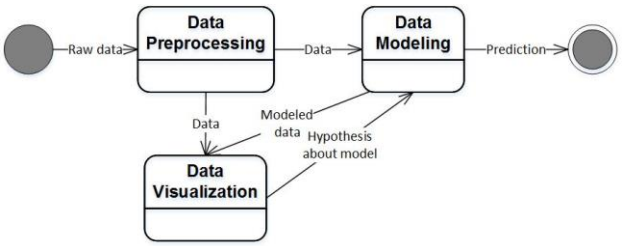

Fig. 7. **Framework methods application for predicting the incidence of COVID-19**
***Source*: Compiled by the author**

The second case is spam filtering described in [27]. The process starts from processing e-mails to encode them in vector description. It continues with clustering, which determines the semantics of similar e-mails, and finishes with classification, supporting the spam filter for incoming e-mail stream. The whole process can be represented as three phases: p_1 – data preprocessing, p_2 – filter modeling, p_3 – spam prediction. The phases are time-, aim- and structure-independent.

Such a process complies with the framework application result. First, data preprocessing is performed. The integrity of the delivered e-mails is the mailing service's responsibility, so there is no need to clean-up the raw data. Therefore, the first task a_1 is data transformation with encoding text letters into a vector representation. The vectors become the input for data modeling. The second task a_2 is a configuration of the spam filter, i.e., to building the filter model. For this purpose, the generative modeling technique allows dividing all letters into two clusters – useful mail and spam. The third task a_3 is to determine the type of incoming e-mails with the spam

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

filter. The appropriate method is predictive modeling; in this case, the classification technique has been used (Fig. 8).

The third case concerned information monitoring and diagnosis of students' performance [28]. The results of each semester controls that are input are clustered. The clusters' sequence corresponded to the results of particular semesters, allowing diagnosing the students' learning behavior.

Following the guiding principles for framework use, the aim can be presented as:

$$A=<a_1, a_2, a_3, a_4, a_5>,$$

where a_1 is a task of data cleansing; a_2 is a task of dimensionality reduction; a_3 is a task of generative modeling of student success; a_4 is a task of getting descriptions of clusters' representatives; a_5 is a task of estimation of students' behavior in the learning process.

Tasks grouping places them in three phases: p_1 – data processing, p_2 – data modeling, p_3 – presentation of diagnostic results.

The aim of the first phase is $A_1=<a_1, a_2>$. For task a_1, the removing observations with missing values did not fit because it caused the loss of information from observations. So all missed values had been replaced with "neutral" values. In the case under consideration, "neutral" is the meaning determined based on the university's internal regulations. For task a_2, the observations were clustered with each subset of features corresponding to a particular semester. The cluster representatives became the input for the next phase. So $MMT_1 = <mmt_{1.1}, mmt_{1.2}>$, $mmt_{1.1}$ is fixing the missing data, $mmt_{1.2}$ is dimensionality reduction.

The aim of the second phase is $A_2=<a_3, a_4>$. Both tasks can be solved with generative modeling, so $MMT_2 = <mmt_2>$, mmt_2 is clustering.

The sequence of applied methods for students' performance diagnostics is shown in Fig. 9.

The fourth case concerned the car sales prediction [29]. The input data was read from the database of cars and commercial vehicle sales per month. An external system collects the data, their quality is guaranteed, so there is no need for preprocessing. However, the authors used visualization to see the analyzed data and modeling properties to obtain descriptive statistics. Predictive modeling was performed by regression analysis realized with a convolutional neural network. Such a choice is based on the trend and seasonality in data and the lack of requirement of a model description of the analyzed data.

Therefore, following the guiding principles for framework use, the aim can be presented as:

$$A=<a_1, a_2, a_3>,$$

where a_1 is a task of exploratory analysis, a_2 is a task of data visualization, a_3 is a task of car sales prediction.

Tasks grouping places them in two phases: p_1 – data exploring, p_2 – predictive modeling.

The sequence of applied methods for car sales prediction is shown in Fig. 10.

The case study approved the consistent of realized solutions with the proposed framework. The framework application is invariant to the problem domain, product development methods, information technologies, etc. That is, the framework is universal.

CONCLUSION

Data science's evolution demonstrates that the three levels hierarchy described in the paper has tended to stability. In the future, it may be useful to add the fourth level, which presents the algorithms. Therefore, the maintenance of framework actuality will not require much effort. For automation of the framework updating the methods of natural language processing will be required.

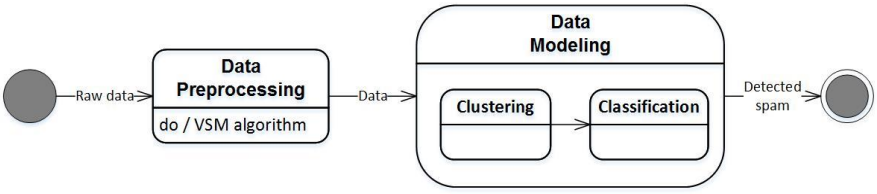

Fig. 8. **Framework methods application for spam filtering**
***Source*: Compiled by the author**

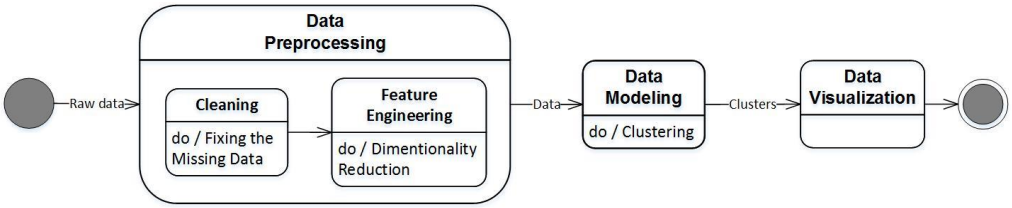

Fig. 9. **Framework methods application for monitoring and diagnostics of students' success**
***Source*: Compiled by the author**

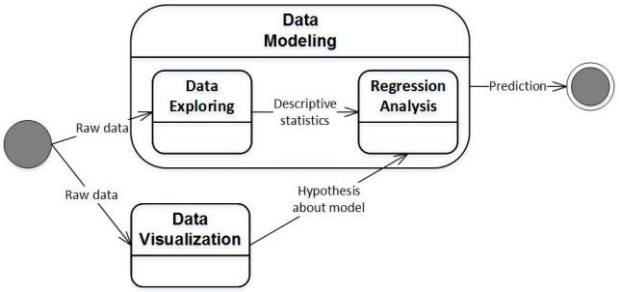

Fig. 10. **Using framework methods to predict car sales**
***Source*: Compiled by the author**

The authors see three areas of the framework application. For data scientists, the framework can help plan work and select relevant technologies to solve the applied problem. For researchers in data science, the framework simplifies the formalization of the appointment of new methods and techniques based on structural relationships. For students studying data science, the framework simplifies finding knowledge gaps and systematizing learned methods and techniques.

The proposed framework is not a sweeping generalization of data science technologies, but it is a basis that can be refined following the purpose of use. The framework is not comprehensive. It covers only those areas where the study objects are described by a set of structured or semi-structured features. Such a restriction, for example, excludes from consideration objects described by texts, as well as the analysis of textual information. However, suppose the textual description of objects is transformed into a description in the space of structured features. In that case, the framework's use is quite possible, which has been demonstrated by the spam filtering described above. The framework also does not cover graph-based network analysis, including social media analysis and audio, video, graphics, and streaming data.

The further work is connected with studying the framework's properties in practice and improving the structure according to the study results. The most interesting is the development of guides and recommendations for the data scientists on the framework using. The first step on this way should be the manual filling the framework with concepts for the various problem domain and problems complexity. In other words, the theoretical model should be projected on the practical problems. The second step should consist of analyzing regularities in the obtained projections based on the logical connections between the framework structural elements and tasks, the probabilities distribution for using certain structural elements of the framework, the detection of anomalous projections, etc. It gives the possibility to develop guides and recommendations for the framework using and support the automation to fill the framework by algorithms further. We also plan to expand the framework by considering natural language processing and network analysis methods.

REFERENCES

1. Vermeulen, A. F. "Practical Data Science: A Guide to Building the Technology Stack for Turning Data Lakes into Business". New York: US: *Apress*. 2018. 805 p. DOI: 10.1007/978-1-4842-3054-1.

2. Rollins, J. B. "Foundational Methodology for Data Science". Somers: US. *IBM Analytics*. 2015. – Available at: https://www.ibm.com/downloads/cas/WKK9DX51. – [Accessed: 17th January 2021].

3. Nasution, M. K. M., Sitompul, O. S. & Nababan, E. B. "Data science". *Journal of Physics Conference Series*. 2020; 1566: 012034. DOI: 10.1088/1742-6596/1566/1/012034.

4. Neifer, T., Lawo, D. & Esau, M. "Data Science Canvas: Evaluation of a Tool to Manage Data Science Projects." *Proceedings of the 54th Hawaii International Conference on System Sciences*. Maui: Hawaii. 2021. p. 5399–5408.

5. Cao, L. "Data Science: A Comprehensive Overview." *ACM Computing Surveys*. 2017; 50(3): 43. DOI: 10.1145/3076253

6. Manovich, L. "Data Science and Digital Art History." *International Journal for Digital Art History*. 2015; 1: 13–35. DOI: 10.11588/dah.2015.1.21631.

7. George, G., Osinga, E. C., Lavie, D. & Scott, B. A. "Big Data and Data Science Methods for Management Research". *Academy of Management Journal*. 2016; 59(5): 1493–1507. DOI: 10.5465/amj.2016.4005.

8. Byeong, S. K., Bong, G. K., Seon, H. C. & Tag, G. K. "Data modeling versus simulation modeling in the big data era: case study of a greenhouse control system". *Simulation: Transactions of the Society for Modeling and Simulation International*. 2017; 93(7): 579–594. DOI: 10.1177/0037549717692866.

9. Akter, Sh. & Wamba, S. F. "Big data analytics in E-commerce: a systematic review and agenda for future research." *Electron Markets*. 2016; 26: 173–194. DOI 10.1007/s12525-016-0219-0.

10. Nguyen, Tr., Zhou, L., Spiegler, V., Ieromonachou, P. & Lin, Y. "Big data analytics in supply chain management: A state-of-the-art literature review." *Research*. 2018; 98: 254–264. DOI: 10.1016/j.cor.2017.07.004.

11. Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martinez-Ortiz, I. & Fernández-Manjón, B.

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

"Applications of data science to game learning analytics data: A systematic literature review". *Computers in Education*. 2019; 141: 103612. DOI: 10.1016/j.compedu.2019.10361.

12. Alonso, S. G., de la Torre Díez, I., Rodrigues, J. J. P. C., Hamrioui, S. & López-Coronado, M. "A systematic review of techniques and sources of big data in the healthcare sector". *Journal of Medical Systems*. 2017; 41(11): 183. DOI: 10.1007/s10916-017-0832-2.

13. Parimbelli, E., Wilk, S., Cornet, R., Sniatala, P., Sniatala, K., Glaser, S., Fraterman, I., Boekhout, A. H, Ottaviano, M. & Peleg, M. "A Review of AI and Data Science Support for Cancer Management". *medRxiv preprint*. medRxiv: 2020.08.07.20170191. 2020. 41 p. DOI: 10.1101/2020.08.07.20170191.

14. Alonso, S. G., de la Torre-Díez, I., Hamrioui, S., López-Coronado, M., Barreno, D. C. & Nozaleda, L. M. "Data mining algorithms and techniques in mental health: a systematic review". *Journal of Medical Systems*. 2018; 42(9): 161. DOI: 10.1007/s10916-018-1018-2.

15. Bao, Y., Chen, Zh., Wei, Sh., Xu, Y., Tang, Zh. & Li, H. "The State of the Art of Data Science and Engineering in Structural Health Monitoring". *Publ. Engineering*. 2019; 5(2): 234–242. DOI: 10.1016/j.eng.2018.11.027.

16. Endert, A., Ribarsky, W., Turkay, C., Wong, B., Nabney, I., Blanco, I. D. & Rossi, F. "The state of the art in integrating machine learning into visual analytics." *Computer Graphics Forum*. 2017; 36(8): 458–486. DOI: 10.1111/cgf.13092.

17. Kemal, M. "Data Visualization Tools In Action Choosing a Visualization Software." *Technical Report*. University of Liverpool. Liverpool: 2019. 11 p. DOI: 10.13140/RG.2.2.11690.26560.

18. Raghav, R. S, Pothula, S., Vengattaraman, T. & Ponnurangam D. "A survey of data visualization tools for analyzing large volume of data in big data platform." *International Conference on Communication and Electronics Systems*. Coimbatore: India. 2016. p. 1–6. DOI: 10.1109/CESYS.2016.7889976.

19. Lowe, J. & Matthee, M. "Requirements of Data Visualization Tools to Analyze Big Data: A Structured Literature Review." *Conference on e-Business, e-Services, and e-Society*. Skukuza: South Africa. 2020. p. 469–480. DOI: 10.1007/978-3-030-44999-5_39.

20. Abdallah, Z. S., Du, L. & Webb, G. I. "Data Preparation". *Encyclopedia of Machine Learning and Data Mining*. Boston: US. *Springer US*. 2017. p. 318–327. DOI: 10.1007/978-1-4899-7687-1_62.

21. Barapatre, D. & A, V. "Data preparation on large datasets for data science". *Asian Journal of Pharmaceutical and Clinical Research*. 2017; 10(13): 485–488. DOI: 10.22159/ajpcr.2017.v10s1.20526.

22. Vyas, S. & Vaishnav, P. "A comparative study of various ETL process and their testing techniques in data warehouse". *Journal of Statistics and Management Systems*. 2017; 20(4): 753–763. DOI: 10.1080/09720510.2017.1395194.

23. Souibgui, M., Atigui, F., Zammali, S., Cherfi, S. & Yahia, S. B. "Data quality in ETL process: A preliminary study". *Procedia Computer Science*. 2019; 159: 676–687. DOI: 10.1016/j.procs.2019.09.223.

24. Wang, J., Wang, X., Yang, Y., Zhang, H. & Fang, B. "A Review of Data Cleaning Methods for Web Information System." *Computers, Materials & Continua*. 2020; 62(3): 1053–1075. DOI: 10.32604/cmc.2020.08675.

25. Flood, R. L. & Carson, E. "Dealing with Complexity. An Introduction to the Theory and Application of Systems Science". New York: US. *Springer US*. 1993. 280 p. DOI: 10.1007/978-1-4757-2235-2.

26. Ayyoubzadeh, S. M., Ayyoubzadeh, S. M., Zahedi, H., Ahmadi, M. & Niakan Kalhori, S. R. "Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study". *JMIR Public Health Surveillance*. 2020; 6(2): e18828. DOI: 10.2196/18828.

27. Bhuiyan, H., Ashiquzzaman, A. & Juthi, T. I. "A Survey of existing E-mail spam filtering methods considering machine-learning techniques." *Global Journal of Computer Science and Technology*. 2018; 18(2): 21–29.

28. Komleva, N., Liubchenko, V. & Zinovatna, S. "Methodology of information monitoring and diagnostics of objects represented by quantitative estimates based on cluster analysis." *Applied Aspects of Information Technology*. *Publ. Science i Technical*. Odesa: Ukraine. 2020; 3(1): 376–392. DOI: 10.15276/aait.01.2020.1

29. Nguyen, Th. Kh. T., Antoshchuk, S. G., Nikolenko, A. A., Tran, K. Th. & Babilunha, O. Yu. "Non-stationary time series prediction using one-dimensional convolutional neural network models." *Herald of Advanced Information Technology*. *Publ. Science i Technical*. Odesa: Ukraine. 2020; 3(1): 362–372. DOI: 10.15276/hait01.2020.3.

Conflicts of Interest: The authors declare no conflict of interest

Received 18.01.2021
Received after revision 02.03.2021
Accepted 15.03.2021

DOI: 10.15276/aait.01.2021.7
УДК 004.02+004.62+004.9

ФРЕЙМВОРК ДЛЯ СИСТЕМАТИЗАЦІЇ МЕТОДІВ НАУКИ ПРО ДАНІ

Віра Вікторівна Любченко[1)]
ORCID: http://orcid.org/0000-0002-4611-7832; lvv@opu.ua
Наталія Олегівна Комлева[1)]
ORCID: http://orcid.org/0000-0001-9627-8530; komleva@opu.ua
Світлана Леонідівна Зіноватна[1)]
ORCID: http://orcid.org/0000-0002-9190-6486; zinovatnaya.svetlana@opu.ua
Катерина Олександрівна Писаренко [1)]
ORCID: https://orcid.org/0000-0001-9573-9315; horodnychaya@opu.ua
[1)] Одеський національний політехнічний університет, проспект Шевченка, 1, Одеса, 65044, Україна

АНОТАЦІЯ

Бурхливий розвиток науки про дані призвів до накопичення великої кількості моделей, методів і технік, які показали доцільність свого застосування. Як показав аналіз публікацій, систематизація методів і технік науки про дані є актуальною задачею, але в більшості результати релевантні до вирішення конкретної прикладної задачі. В роботі розроблено не орієнтований на конкретну предметну область та задачу фреймворк для систематизації методів науки про дані. Трирівнева ієрархія метамодель-метод-техніка впорядковує залежності між існуючими методами та техніками та знижує складність їх розуміння. Перший рівень структури складають три метамоделі: препроцесингу, моделювання та візуалізації даних. На другому рівні розташовані методи, які відповідають метамоделям. На третьому рівні зібрані основні техніки, згруповані відповідно до методів. Також описано принципи використання ієрархії. Для цього формально визначено типовий процес вирішення завдання з залученням методів науки про дані та сформульовано алгоритм використання фреймворку. Метод ситуаційного аналізу застосовано для того, щоб пересвідчитися в працездатності фреймворку. Розглянуто чотири кейси застосування методів науки про дані для вирішення практичних завдань, які описані в сучасних публікаціях. Показано, що описані рішення повністю узгоджуються з запропонованою ієрархією методів фреймворку. Визначено рекомендовані напрямки застосування фреймворку та його обмеження: об'єкти обраної галузі повинні бути описані лише структурованими або напівструктурованими ознаками. На завершення наведено шляхи подальших досліджень.

Ключові слова: наука про дані; фреймворк; попередня обробка даних; моделювання даних; візуалізація даних; ситуаційний аналіз.

ABOUT THE AUTHOR

Vira Liubchenko – Dr. Sci. (Eng.) (2014), Cand. Sci. (Eng.) (1997), Professor, Department of System Software, Odesa National Polytechnic University, 1, Shevchenko Ave, Odesa, 65044, Ukraine
ORCID: http://orcid.org/0000-0002-4611-7832; lvv@opu.ua
Research field: Data Science, Software Engineering, Project Management

Віра Вікторівна Любченко – д-р техніч. наук (2014), канд. техніч. наук (1997), професор кафедри Системного програмного забезпечення Одеського національного політехнічного університету, проспект Шевченка, 1, Одеса, 65044, Україна

Nataliia Komleva – Cand. Sci. (Eng.) (2006), Associate Prof., Department of System Software, Odesa National Polytechnic University, 1, Shevchenko Ave, Odesa, 65044, Ukraine
ORCID: http://orcid.org/0000-0001-9627-8530; komleva@opu.ua
Research field: Data Analysis, Software Engineering, Knowledge Management

Наталія Олегівна Комлева – канд. техніч. наук (2006), доцент кафедри Системного програмного забезпечення Одеського національного політехнічного університету, проспект Шевченка, 1, Одеса, 65044, Україна

Svitlana Zinovatna – Cand. Sci. (Eng.) (2008), Associate Prof., Department of System Software, Odesa National Polytechnic University, 1, Shevchenko Ave, Odesa, 65044, Ukraine
ORCID: http://orcid.org/0000-0002-9190-6486; zinovatnaya.svetlana@opu.ua
Research field: Data Analysis, Information System Productivity

Світлана Леонідівна Зіноватна – канд. техніч. наук (2008), доцент кафедри Системного програмного забезпечення Одеського національного політехнічного університету, проспект Шевченка, 1, Одеса, 65044, Україна

Katherine Pysarenko – Cand. Sci. (Eng.) (2018), Associate Prof., Department of System Software, Odesa National Polytechnic University, 1, Shevchenko Ave, Odesa, 65044, Ukraine
ORCID: https://orcid.org/0000-0001-9573-9315; horodnychaya@opu.ua
Research field: Data Analysis, Software Engineering

Катерина Олександрівна Писаренко – канд.техніч.наук (2018), доцент кафедри Системного програмного забезпечення Одеського національного політехнічного університету, проспект Шевченка, 1, Одеса, 65044, Україна

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)