https://doi.org/10.21122/1029-7448-2026-69-1-5-22

UDC 004.048; 004.051

Improving Performance of Fitting Energy-Saving Measures of Buildings Using Artificial Neural Networks

V. K. Stan[1], D. G. Bukhanov[1], Yu. A. Koshlich[1]

[1]Belgorod State Technological University named after V. G. Shukhov
(Belgorod, Russian Federation))

Abstract. The article is devoted to increasing energy saving and energy efficiency of public sector institutions by automation of the process of energy-saving measures fitting. A classifier based on an artificial neural network is proposed as a technology for selecting measures. The base numeric dataset was supplemented by categorical data on buildings and structures. The urgency of the work is justified by the need to develop solutions aimed at introducing measures to ensure energy saving and energy efficiency in the public sector. The structure and operating principle of a module of auto-selecting energy saving measures as a part of the Energy Resources Management System (SUER) are described. The work studies the quality of fitting energy-saving measures with or without using data on categorical features. An analysis of the selected categorical features, as well as a comparative analysis of their encoding methods, is carried out. An accuracy evaluation method of the classifier in the context of the work is proposed. A series of experiments were conducted by enumerating combinations of numeric and categorical data converting methods, in order to determine the significance of categorical features in general, as well as to determine the most effective combination of encoding methods. A comparative analysis of the results of the experiments was performed and the most successful model was determined, with an additionnal assessment of the quality of the model was made based on the metrics of precision, recall and F1-score. Conclusions were made on the advisability of supplementing the original dataset with categorical data to improve the performance of the system.

Keywords: artificial neural networks, energy saving programs, energy audit, categorical data

For citation: Stan V. K., Bukhanov D. G., Koshlich Yu. A. (2026) Improving Performance of Fitting Energy-Saving Measures of Buildings Using Artificial Neural Networks. *Energetika. Proc. CIS Higher Educ. Inst. and Power Eng. Assoc.* 69 (1), 5–22. https://doi.org/10.21122/1029-7448-2026-69-1-5-22

Повышение качества подбора энергосберегающих мероприятий зданий при использовании искусственных нейронных сетей

В. К. Стан[1], Д. Г. Буханов[1], Ю. А. Кошлич[1]

[1]Белгородский государственный технологический университет имени В. Г. Шухова
(Белгород, Российская Федерация)

Реферат. Статья посвящена повышению энергосбережения и энергоэффективности учреждений бюджетной сферы за счет автоматизации процесса подбора энергосберегающих

Адрес для переписки	**Address for correspondence**
Стан Василий Константинович	Stan Vasily K.
Белгородский государственный технологический	Belgorod State Technological
университет имени В. Г. Шухова	University named after V. G. Shukhov
ул. Костюкова, 46,	46, Kostyukova str.,
308012, г. Белгород, Российская Федерация	308012, Belgorod, Russian Federation
Тел.: +7 (4722) 30-99-64	Tel.: +7 (4722) 30-99-64
madseal@yandex.ru	madseal@yandex.ru

мероприятий. В качестве технологии подбора мероприятий в работе предложен классификатор, построенный на базе искусственной нейронной сети. Набор информационных признаков дополнен категориальными данными зданий учреждений. Актуальность работы обоснована необходимостью разработки решений, направленных на внедрение мер по повышению энергосбережения и энергоэффективности в бюджетной сфере. Описаны структура и принцип работы модуля автоматического подбора мероприятий по энергосбережению в составе системы управления энергетическими ресурсами (СУЭР). В работе произведены исследования качества подбора энергосберегающих мероприятий с наличием и отсутствием категориальных признаков. Произведен анализ выделенных категориальных признаков, а также сравнительный анализ методов их кодирования. Предложена методика оценки точности работы классификатора в контексте решаемой задачи. Путем перебора комбинаций методов преобразования количественных и категориальных данных проведен ряд экспериментов с целью определения значимости категориальных признаков в целом, а также определения наиболее результативного сочетания методов их кодирования. Произведен сравнительный анализ полученных результатов с определением наиболее успешной модели, для которой была произведена дополнительная оценка качества работы на базе метрик точности, отклика и средневзвешенной F-меры. Сделаны выводы о целесообразности дополнения исходного набора категориальными данными для улучшения показателей работы разработанной системы.

Ключевые слова: искусственные нейронные сети, программы энергосбережения, энергоаудит, категориальные данные

Для цитирования: Стан, В. К. Повышение качества подбора энергосберегающих мероприятий зданий при использовании искусственных нейронных сетей / В. К. Стан, Д. Г. Буханов, Ю. А. Кошлич // *Энергетика. Изв. высш. учеб. заведений и энерг. объединений СНГ*. 2026. Т. 69, № 1. С. 5–22. https://doi.org/10.21122/1029-7448-2026-69-1-5-22

Introduction

A necessary condition for the development of any state is continuous economic growth, accompanied by an increasing in the consumption of resources and its consumers [1–4]. The main types of energy resources consumed by all sectors of the economy are heat and electrical energy, cold and hot-water supply, natural gas, etc. At the same time, the availability of resources in the public sector, critically important and sensitive to the level of resource provision, is affected by both the availability of the resource itself and the features of the financial policy being pursued [5–7]. Federal Law No 261-FZ “On Energy Saving and Improving Energy Efficiency, and on Amendments to Certain Legislative Acts of the Russian Federation”, adopted in 2009, is aimed at “creating a legal, economic and organizational framework for stimulating energy saving and improving energy efficiency” [8]. An important addition to this law is the fact that the Russian government has developed several energy strategies up to 2035 and 2050 [4].

In practice, the achievement of the set goals is ensured by conducting an energy audit of the facility with the subsequent creation and control of an energy saving program (ESP). In this case, the basis of the ESP is a set of measures, developed within the framework of the energy audit and aimed at achieving target energy saving indicators [9–11].

The study of the subject area showed that the most difficult stage of the process is the energy audit due to economic and project risks. On the one hand, the costs of conducting research and implementing energy saving measures (ESMs) may not be recouped within the expected useful life of a building. On the other hand, classical methods of conducting energy audit do not cover all the typological groups of buildings and measures, and not excluding influence of the specialists' competencies involved in the work [12, 13].

The problem of reduction the cost of money and human resources for conducting an energy audit and the effectiveness of the proposed energy-saving measures arises. Thus, the goal of the work is to automate the fitting of ESMs. To solve the problem, a number of software systems based on artificial intelligence methods: data clustering, fuzzy logic, expert systems, genetic algorithms and methods based on the use of artificial neural networks (ANNs) have been developed. The analysis of this systems showed that there is no common solution, but at the same time, a large number of them are proposed for individual types of buildings. Also, most of the solutions considered require high resource costs for preparing data or employees to work with them [14, 15]. In view of the wide development and areas of application, when solving such problems, the work proposes to use ANN [16].

As a result, within the framework of the Energy Resources Management System (SUER) [17], an intelligent module, that performs the selection of energy-saving measures for a building within the framework of the creation of an ESP, was developed. The structure of the module is shown in Fig. 1.

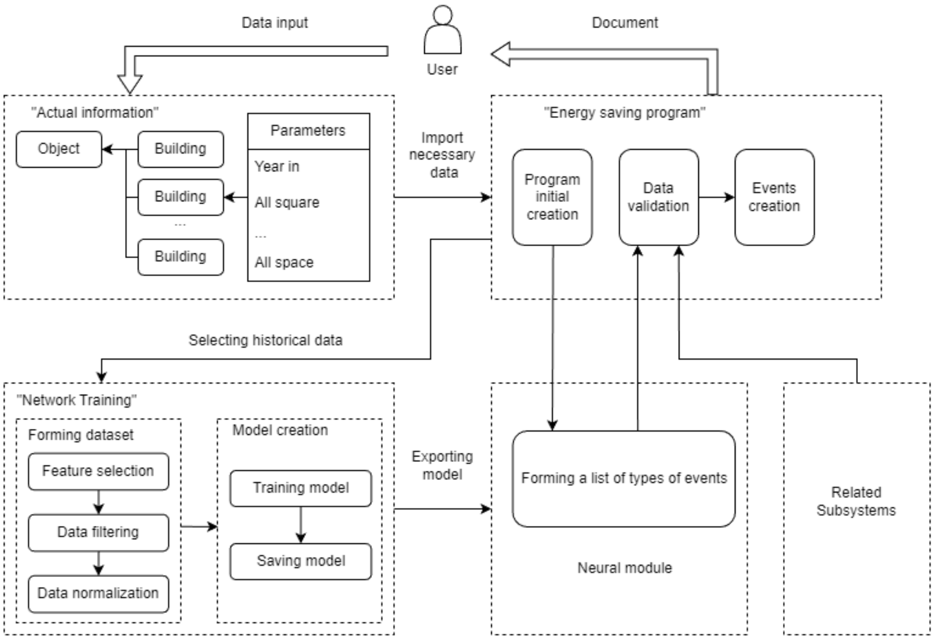

Fig. 1. Diagram of the process of operation of the subsystem of automatic selection of events of the SUER

At the initial stage, the user [*User*] enters data about institutions [*Data Input*] into the *"Actual information"* subsystem [*Actual Information*]. This data represents essential information about the institution [*Object*], as well as information on the buildings included in its structure [*Building*]. Buildings have a number of characteristics [*Parameters*], which make it possible to, directly or indirectly, evaluate them. Next, at the initial stage of the ESP generation, all the necessary information is imported from *"Actual Information"* [*Import Necessary Data*] into *"Energy Saving Programs"* [*Energy Saving Programs*]. After, a program with a basic level of filling will be created [*Program Initial Creation*]. The next stage is the generating of a list of events [*Forming a list of types of events*]: the essential information is transformed and transmitted to the input of the neural network [*Neural Module*]; the network processes the information and returns at the output a list of event types, which represents a vector of the probabilities of using a particular ESM. The received information is converted to a normal form and then undergoes verification [*Data Validation*] for the possibility of using a particular event based on data from other subsystems [*Related Subsystems*]. At the final stage, events are created and appraised [*Events Creation*] and a ready to print ESP [*Document*] can downloaded by *User* [*Getting document*].

The neural network fitting process [*Network training*] should be highlighted separately. At the preliminary stage, a data sample [*Forming Dataset*] is dumped from the ESP subsystem [*Selecting Historical Data*]. This sample undergoes filtration procedures [*Data Filtering*] and normalization [*Data Normalization*] of a pre-selected set of building parameters [*Feature Selection*]. The resulting dataset is used to fit the network [*Training Model*], the result of which is a file containing its parameters and structure [*Saving Model*]. This process is continuous, since the database is constantly updated with new ESPs.

The data, which represent information on previously developed ESPs and are accumulated during the operation of the SUER, were used to train the ANN. Due to the architecture of the system it was decided to use information on the building as a part of the institution as a input data vector. At the stage of the initial formation of the sample, the following features were excluded from the general set:

– *categorical features*, since it is difficult to assess their influence within a feature on the output data and the work of the classifier: type of a building and institution, etc.;

– *numeric features* that are optional and have a low occupancy rate: width and length of a building.

Thus, a sample was formed that included about 12,000 records containing 22 numeric features at the input:

– *volumes of energy resource use:* volumes of energy resources consumed in the year preceding the year in the ESP started;

– *structural characteristics of buildings:* year built, total volume and area, useful and heated volumes and area, areas of dispersion surfaces (outwalls, attic and basement), number of windows, etc.;

– *operational characteristics of buildings:* number of employees, number of visitors.

And one output feature – a list of energy saving event type identifiers, converted using Multi-Hot Encoding.

To improve the quality of training and subsequent operation of the network, the data were cleaned and normalized. Data cleaning from outliers was performed using the Box-plot method [18] and a priori information: outliers were identified based on the specificity of a particular feature using visual classification. The result of data cleaning was the exclusion of about 2000 records from the total sample, or 17 % of the original dataset.

Data normalization was performed using min-max normalization (scaling): the result of this type of normalization is bringing the feature values into the range [0; 1] (Table 1). In the context of this work, the need to use this operation is justified by the large spread of the initial data: the values in the final sample lie in the range from 0 to $5 \cdot 10^5$.

Two methods for data normalization were chosen: linear and logarithmic scaling [19]. Linear scaling has the following form:

$$y = \left(x - \min\left(x\right)\right) / \left(\max\left(x\right) - \min\left(x\right)\right), \tag{1}$$

where y – normalized data; x – source data; max(x) and min(x) – maximum and minimum values in the sample.

Logarithmic normalization has the following form:

$$y = x - \min\left(x\right) / 10^j, \tag{2}$$

where y – normalized data; x – source data; min(x) – minimum values in the sample; j – scaling factor,

$$j = \log_{10}\left|\max\left(x\right) - \min\left(x\right)\right|. \tag{3}$$

Table 1

Example of normalized data for training a neural network

Year in	Heat consumption	Water consumption	…	All space	Windows count	Event ids
0.622951	0.025641	0.045384	…	0.042267	0.086957	4, 38, 40
0.467213	0.013704	0.002771	…	0.045721	0.060201	36, 38, 40, 41, 50, 51
…	…	…	…	…	…	…
0.54918	0.035985	0.101683	…	0.097881	0.197324	38, 41

Based on the analysis of some papers [20–23], a classifier based on feedforward neural network (multilayer perceptron) was developed. The choice of this type of network is due to its applicability in solving this type of problems (multilabel classification).

As a result, a neural network with the structure shown in Fig. 2 was developed [24, 25].

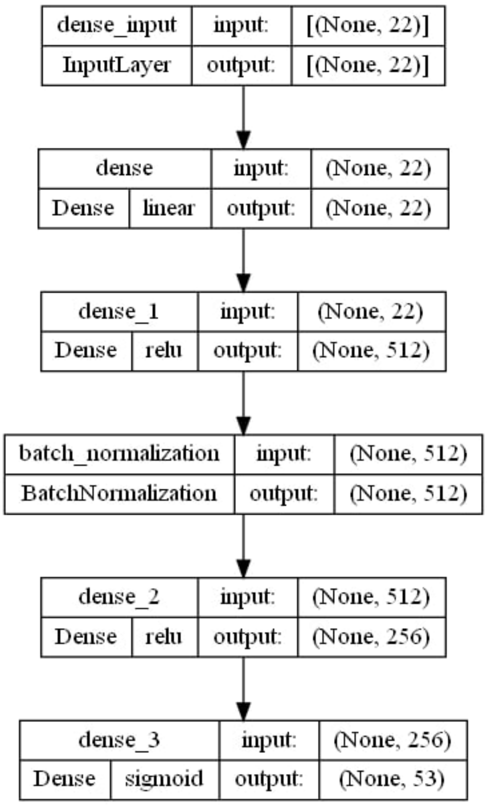

Fig. 2. Structure of the original neural network

The network implementation is done in Python 3.10 using Keras API (v2.13.1) on top of TensorFlow framework (v2.13.1), trained on Intel Core i7-12700 CPU with DDR5 SDRAM and has the following configuration code:

```
model = M.Sequential()
model.add(L.Dense(inputs, input_dim=22, activation='linear'))
model.add(L.Dense(512, activation='relu'))
model.add(L.BatchNormalization())
model.add(L.Dense(256, activation='relu'))
model.add(L.Dense(53, activation='sigmoid'))
model.compile(optimizer="adam", loss='mean_squared_error', metrics='accuracy').
```

The basic metrics for evaluating the results of classifier training are *accuracy* and *loss* function.

Accuracy is calculated as the ratio of correctly predicted values to the total number of iterations.

Loss function (error): mean squared error (MSE), calculated by the formula (4), is used:

$$MSE = \frac{1}{n}\sum\left(y - \hat{y}\right)^2, \tag{4}$$

where $\hat{y}$ is the predicted value, y is the real output value, n is the total number of training operations.

The results of training the developed model on numeric data using the prepared dataset are shown in Fig. 3–4.

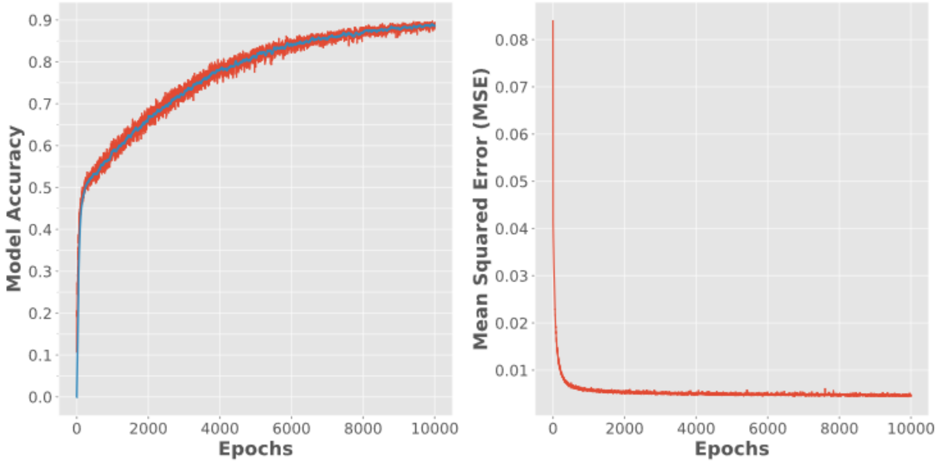

Fig. 3. Classification quality using linear scaling

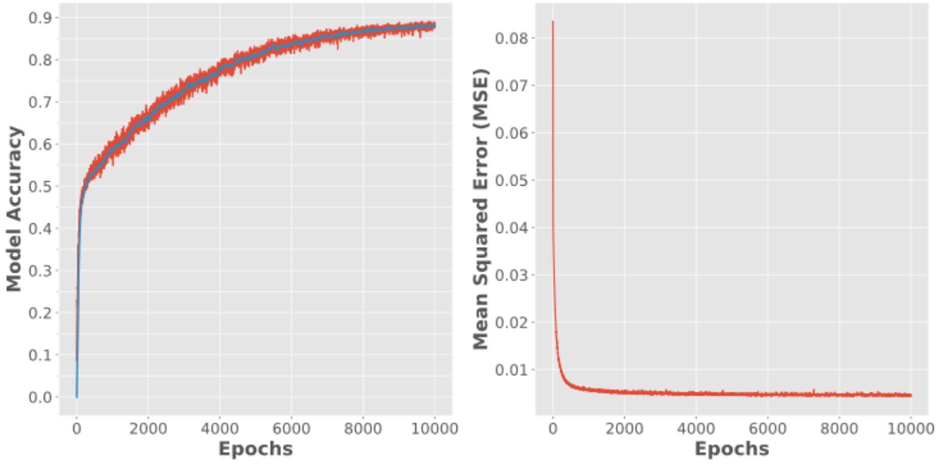

Fig. 4. Classification quality using logarithmic scaling

An assessment of the obtained results showed that the proposed approach proved its feasibility and effectiveness. However, the previously imposed restrictions on the data used in the classifier's work could have caused the lower accuracy of the network. To improve the accuracy, it is necessary to add an analysis of categorical features. This requires an assessment of the structure and quality of the data, determining their correlation with the output and numeric data, and determining the most effective method for their normalization.

Preparing for experiments

To improve the quality of classification, it was decided to supplement the dataset with categorical features. Thus, 12 categorical features were added to the 22 numeric ones (Table 2). They contain information about the functional purpose of buildings, the presence of wall insulation, glazing area and a number of other characteristics that directly affect the energy saving characteristics of buildings, as well as the ways to improve them.

The categorical data block can be structurally divided as follows:

– *Belonging to typological groups:* type of building; type of institution in whose structure the building is included; energy efficiency class of the building; functional purpose of the building;

– *Design features:* type of attic floors and the method of their insulation; material of window units and their execution; material from which the walls are made, type of basement rooms;

– *Flags (binary features):* technical floor, presence of facade insulation.

Table 2

Source categorical dataset sample

No	Type organization	Type building	Kee	Ext wall material	Type of building	Façade heat having	Wood window block type	Plastic window block type	Roof	Warm roof	Cellar
1	2	8	4	False	2	False	1	3	3	1	1
2	1	9	4	False	4	True	1	3	3	6	3
3	2	8	1	False	2	False	1	2	3	1	1
4	11	3	8	False	2	False	1	2	2	1	1
5	1	9	3	False	4	True	1	2	3	7	1

From the number of unique values corresponding to each of the features (Fig. 5), it is clear that the minimum number per feature was two (binary features), and the maximum was 24 (for "Type building").

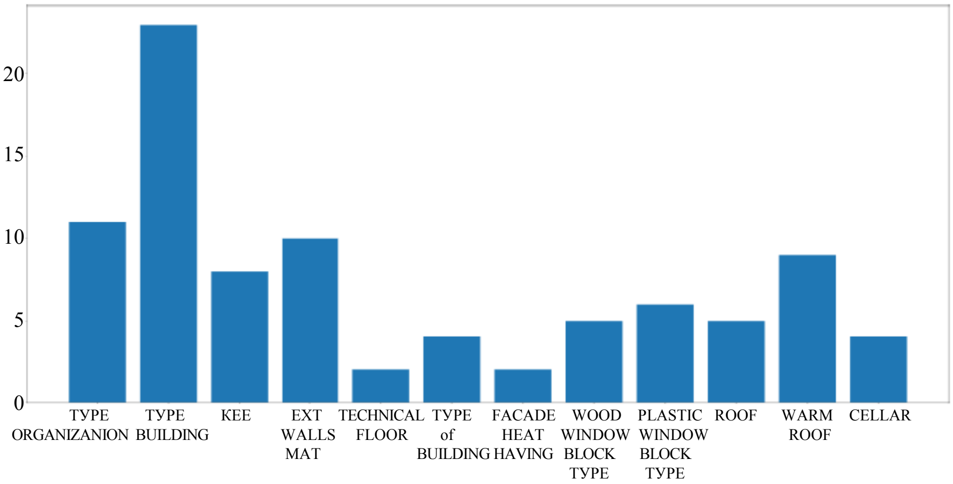

Fig. 5. Number of unique values of categorical features:
X-axis – categorical features; *Y*-axis – number of unique values for each feature

The specifics of the data used are also determined by the method of storing them in the database. In this case, each category is already assigned a serial number (often from 1 to *N*). Thus, it can be considered that the categorical data describing typological groups of building features are initially transformed using Label Encoding [26], which will be described below.

To improve the quality of the initial data for analysis, it was decided to evaluate the possibility of switching from Label Encoding to Ordinal Encoding. For this purpose, graphs of the distribution of values within each of the features were obtained (Fig. 6). Using the example of the *Type Building* feature, it can be seen that the order number in the initial data does not in any way affect the occurrence of the value in the dataset.

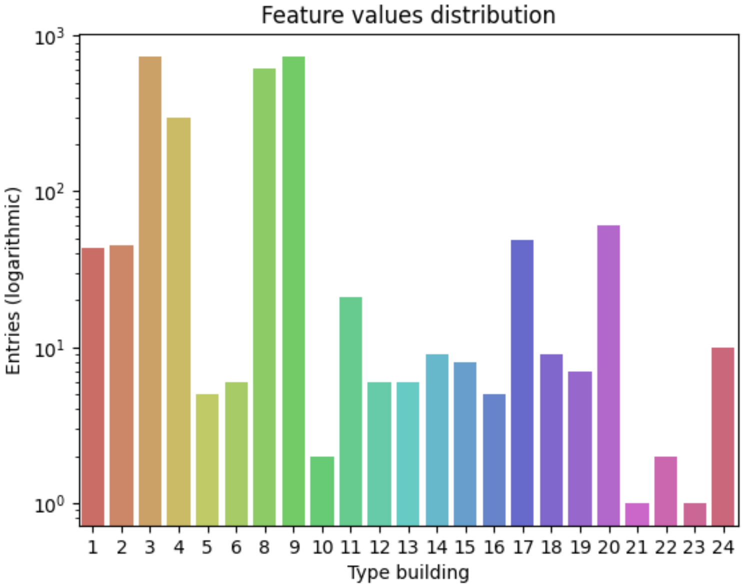

Fig. 6. Distribution of values using an example of "Type building" feature: *X*-axis – serial number of a "Type building" unit; *Y*-axis – number of entries in logarithmic scale

Categorical data encoding

To work with categorical data, it is necessary to select the type of their transformation. There is a wide range of methods for encoding categorical data depending on the task being solved. All these methods are similar in the need to determine the set of values corresponding to the encoded feature. Let us consider the most common ones taking into account the specifics of the coded data [26–34].

One-Hot Encoding: the value of the target feature is replaced by a vector consisting of a set of zeros (0) and only one "one" (1), the position of which corresponds to the position of the feature in the set of unique values [26, 28]. There is a variation of this method called *Multi-Hot Encoding*, which allows storing the value of several features at once and is used to transform output data: a set of integer identifiers of energy saving measures is replaced by a set of ones located in the vector at indices equal to the value of the type. This method allows encoding any data regardless of its type with the possibility of reverse encoding, but the number of features at the input increases in direct proportion to the number of values it receives.

Label Encoding: when using this method, each unique value of the feature is assigned an integer (sequence number). The main disadvantage of this method is the creation of redundant dependencies in the data, since from the point of view of the neural network, the place in the list determines the significance of the feature value. Most of the existing categorical features in source dataset have already encoded with it due to the specifics of storing them in the database. There is a variation of the method called *Ordinal Encoding* [26, 27, 29], the main difference of which is the compilation of a correspondence table for encoding features. Using of this approach is suitable only for ordered series, like scales of subjective assessments, and it allows you to compensate the disadvantages of Label Encoding in some degree.

Binary Encoding is a hybrid of *Label* and *One-Hot Encoding*: each of the N-unique values of a feature is assigned with an integer index, which in turn is entered into a set of binary features (just a binary representation of the labels) [26]. When using this method, the growth of the final number of features corresponds to the number of digits of the maximum value of the feature and changes according to log(N). However, while compensating for some of the disadvantages of the original methods, it inherits a number of other ones: redundancy of encoding and the creation of redundant dependencies in the data.

Frequency Encoding: it is a simple method, since the share (occurrence) of the encoded value in the dataset replaces the value of the feature in it. The advantages include the efficiency of encoding, data scalability, since the encoded values are in the range from 0 to 1, and the introduction of a dependence of the significance of the feature on the frequency of its use.

Mean Encoding / Target Encoding involves transforming a feature in accordance with the value of the target feature [26, 29, 30]. It also depends on the classification task: for regression – the average value of the target label for a given value of the encoded feature; for binary classification – the probability of a single class (event occurrence) for a given value of the encoded feature. There is a simplified version called *M-estimator Encoding* that uses a scaling factor to transform values. It also has varieties like *Leave One Out Encoding* and *CatBoost Encoding*, in which the current vector is excluded when calculating the value of the target feature. This family of methods also includes *James-Stein Encoding, Weight of Evidence Encoding* and its variation – *Probability Ratio Encoding*, based on comparing the encoded values with the average values of the target feature. In case of the problem being solved, using these methods have no clue, since the target feature is a set of values and encoding them into one contradicts the final goals of the multi-class classification being produced.

Hashing Encoding is similar to *Binary Encoding*, except for the fact that the number of columns (digits) is determined not by the number of unique values of the feature, but manually (usually from 1 to 100). It suits well for encoding text data, but inappropriate in this case because the encoded features are already integer.

In addition, since the original categorical data are integers, the transformation methods previously used for numeric data (linear and logarithmic scaling) are applicable to them. It will allow to fit the data into the previously established interval [0; 1].

Ultimately, to find a solution to the problem, several methods of transforming categorical data were selected and the following experiments were conducted:

– solving a classification problem using only categorical data without any transformation and scaling to determine their value through assessing the correlation with the output;

– solving a classification problem using scaled numeric data and categorical data without transformations to assess the impact of categorical data on the accuracy of the classification process in general;

– solving a classification problem using scaled numeric data and scaled categorical data to assess the potential impact of encoding methods on classification accuracy;

– solving a classification problem using scaled numeric data and categorical data transformed using the following methods: Binary Encoding, Frequency Encoding, Helmert Encoding, Backward-Difference Encoding to identify the most effective encoding method within the framework of solving the problem.

Conducting experiments

To conduct the noted experiments, the number of perceptrons in the hidden layers was increased from 512/256 to 700/400, respectively. This was done because the graphs showed underfitting of the resulting neural network when the number of input parameters increased.

A metric for testing the quality of the model was also developed, because the standard accuracy metric used in Keras API [35] calculates accuracy by counting only full matches of values, but given the fact that the output data is an array of event types, this technique is redundant. The evaluation of the results using the developed technique is carried out in two stages:

1. Since the result of the neural network is a vector of probability estimates, the data is transformed into an “encoded” form by setting a threshold value (0.8) to assess if an ESM fits. Then the resulting set of zeros and ones is transformed into an integer series of identifiers in order to discard the comparison of missing events (0 when encoding). Thus, for example, the output vector (1.311^{-31}, 2.764^{-17}, 1.0005^{-18}, 7.953^{-9}, 4.298^{-13}, 2.43^{-34}, 3.798^{-37}, 1.826^{-16}, 0.0, 1.795^{-29}, 2.504^{-22}, 2.276^{-17}, 0.0, 3.27^{-32}, 6.41^{-22}, 1.376^{-17}, 3.454^{-33}, 7.404^{-34}, 1.076^{-35}, 9.33^{-29}, 6.858^{-17}, 1.453^{-14}, 1.309^{-14},1.05^{-15}, 8.145^{-14}, 3.001^{-15}, 1.759^{-15}, 3.874^{-21}, 2.127^{-11}, 9.999^{-1}, 9.057^{-19}, 4.582^{-21}, 3.556^{-22}, 9.154^{-10}, 7.732^{-34}, 1.0, 0.0, 7.504^{-8}, 3.748^{-6}, 1.0, 7.424^{-8}, 9.99^{-1}, 5.946^{-27}, 8.385^{-11}, 1.196^{-31}, 1.041^{-10}, 1.602^{-29}, 5.995^{-14}, 1.584^{-16}, 2.241^{-13}, 4.373^{-11}, 3.383^{-33}, 5.037^{-29}) will first take the form (0000000000 0000000000 0000000001 0000010001 0100000000 0000000000), and at the final stage it gets form like (30, 36, 40, 42);

2. The accuracy of the prediction is assessed by comparing the initial set of identifiers with those obtained as a result of the neural network: the prediction is considered successful if more than 50% of the events in the initial set are included in the predicted set (for extra events, it is assumed that they will be filtered out at the validation stage in the “*Data Validation*” module).

The developed metric uses a 0.8 threshold value. This number was selected empirically based on available statistical data.

To evaluate the accuracy of the work, a test data set was prepared with a number of rows corresponding to 10 % of the original dataset.

Results

As a result, a number of experiments were conducted:

Experiment 1: Using only categorical data for fitting without any transformations. The resulting graphs (Fig. 7) show a correlation between the input and output data, but it is not as strong as in the case of numeric data.

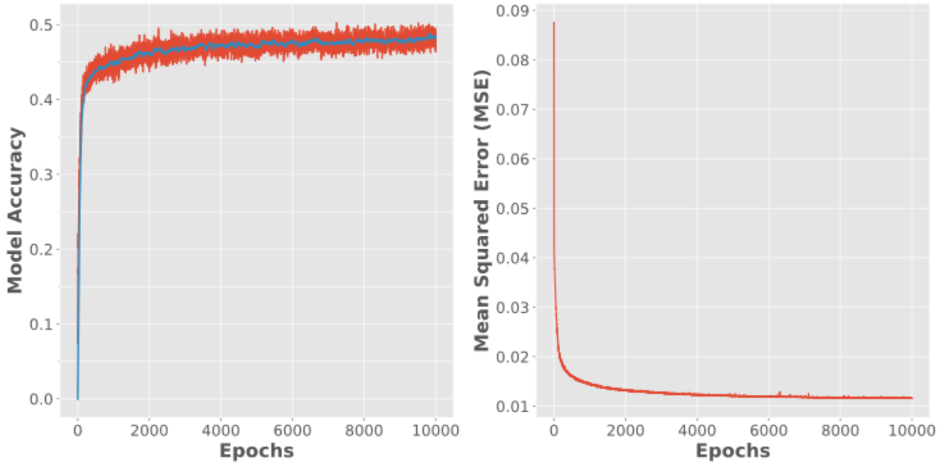

Fig. 7. Metrics of ANN fitting using only categorical data without any transformations

Experiment 2: Fitting a network using linearly scaled categorical data only (Fig. 8). The data obtained show that reducing the values to a single interval does not affect the fitting quality. The data obtained may also indicate that the serial number of categorical data unit does not correlate with its impact.

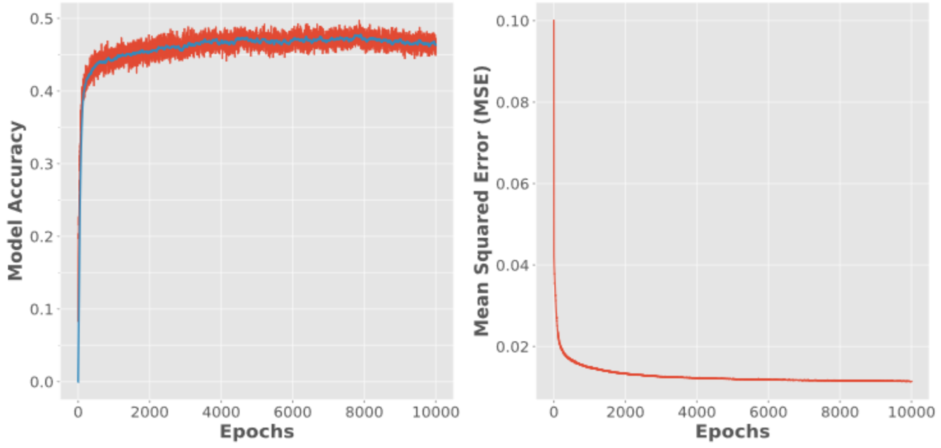

Fig. 8. Metrics of ANN fitting using transformed with linear scaling categorical data only

Experiment 3: Network fitting using numeric data with scaling and categorical data without any transformations (Fig. 9–10). Based on the results obtained, it can be concluded that the introduction of categorical data into the set, although slightly, increased the accuracy of classification.

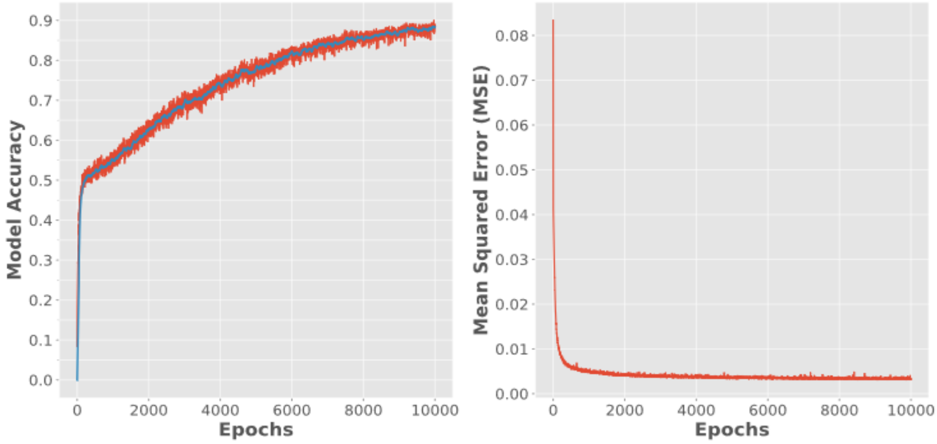

Fig. 9. Metrics of ANN fitting using linearly scaled numeric data and categorical data without any transformations

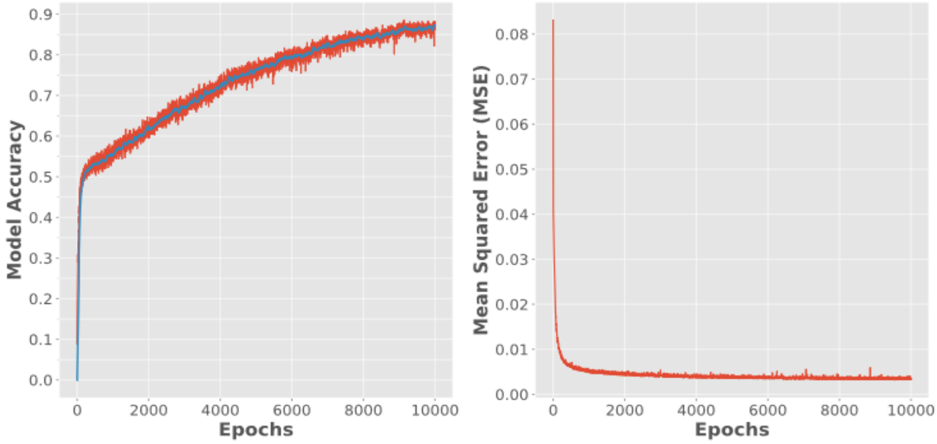

Fig. 10. Metrics of ANN fitting using logarithmically scaled numeric data and categorical data without any transformations

Experiment 4: Fitting ANN using scaled numeric data and linearly scaled categorical data (Fig. 11–12). In this case, higher classification accuracy rates are observed comparing to the previous experiment, which is probably due to the reduction of all values to a single range [0; 1].

Experiment 5: Network fitting using scaled numeric data and categorical data transformed using frequency encoding (Fig. 13–14). This combination of data transformation methods allowed to achieve fitting accuracy above 0.91 in both cases, which can be considered an improvement in compare with fitting using numeric data only.

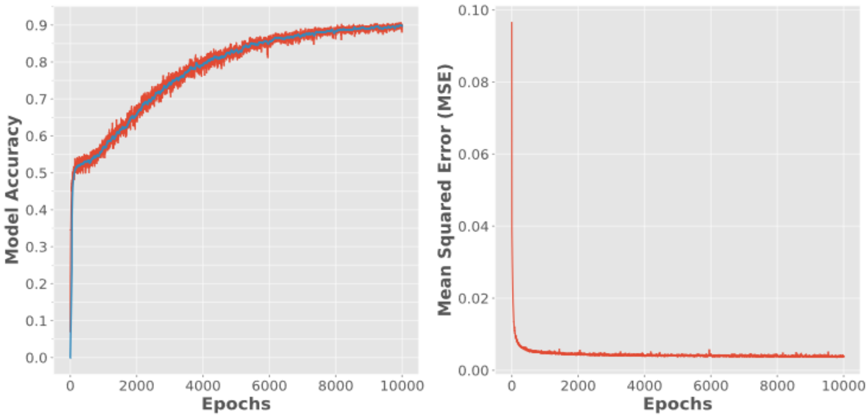

Fig. 11. Metrics of ANN fitting using linearly scaled numeric data and linearly scaled categorical data

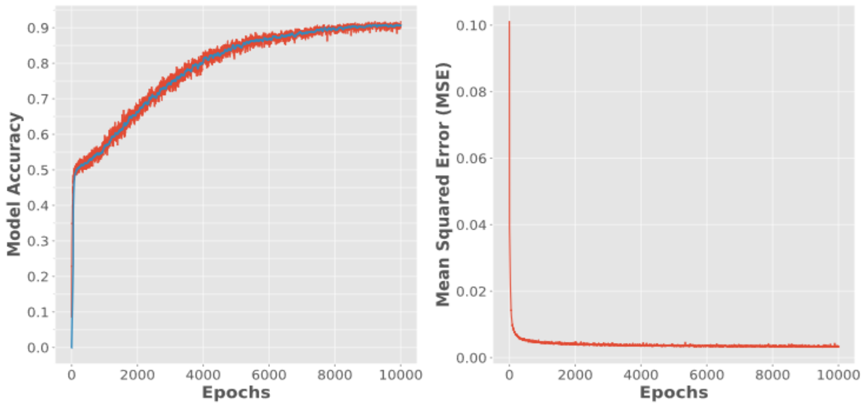

Fig. 12. Metrics of ANN fitting using logarithmically scaled numeric data and linearly scaled categorical data

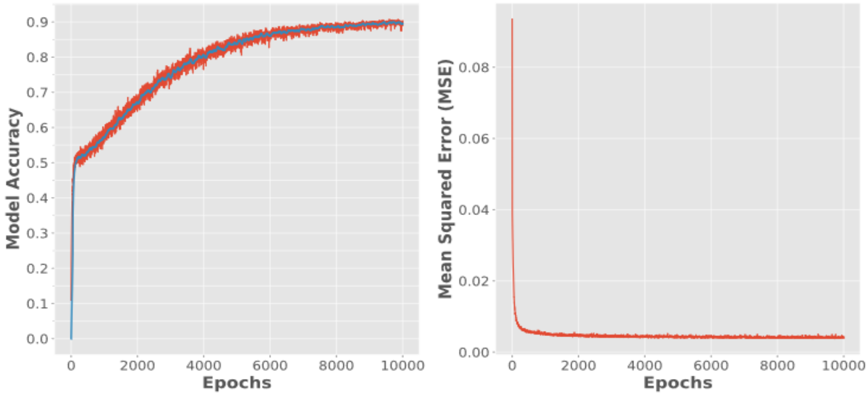

Fig. 13. Metrics of ANN fitting using linearly scaled numeric data and categorical data transformed using frequency encoding

In the case of experiments using Helmert, Backward-Difference and Binary Encoding methods, the fitting ANN using the initial model configuration failed. Attempting to correct the situation by changing the model hyperparameters (its architecture, layer structure and fitting setup), as well as additional data preparation methods, did not help to fix it.

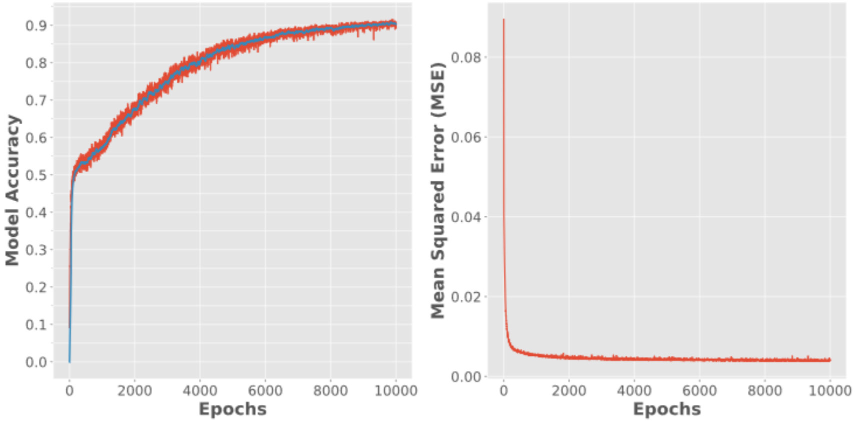

Fig. 14. Metrics of ANN fitting using logarithmically scaled numeric data and categorical data transformed using frequency encoding

The summary of the results of fitting and testing the obtained models is presented in Table 3.

Table 3

Quantitative indicators of the results of training and testing the model based on different data configurations

No	Numeric data encoding	Categorical data encoding	Model accuracy (Fitting)	Model accuracy (Testing)
1	Linear scaling	Data excluded	0.8836	0.89
2	Logarithmic scaling	Data excluded	0.872	0.86
3	Data excluded	Without converting	0.4833	0.65
4	Data excluded	Linear scaling	0.4751	0.53
5	Linear scaling	Linear scaling	0.9065	0.46
6	Logarithmic scaling	Linear scaling	0.9124	0.43
7	Linear scaling	Frequency encoding	0.9058	0.91
8	Logarithmic scaling	Frequency encoding	0.902	0.87
9	Linear scaling	Binary, Helmert, Backward-Difference Encoding	0.0051	0.0

Discussion

The analysis of the obtained results allows us to consider that categorical data have value and using it in the process of forming a set of ESMs is justified. At the same time, the significance of this data class, compared to numeric data, is significantly lower, and it is demonstrated by the accuracy of the fitting models using only numeric data (89 %) and categorical data only (65 %). It is clear, that the methods previously used to transform numeric data are not applicable to categorical data: the use of scaling on isolated categorical data did not have any effect, and when implemented, fitting on a combined dataset even reduced the accuracy of the model from 89 to 46 %.

A practical comparison of the results of using categorical data encoding methods showed that the frequency encoding method is the most effective.

At the same time, some of methods that use dummy variables turned out to be inapplicable for fitting current configuration of the classifier. On the other hand, the results of works [28, 29], show a relatively high efficiency of methods using dummy variables (One-Hot, Helmert, etc.) in solving linear regression problems. It is worth noting that these works are similar by the fact that all surveys were carried out for problems of single-class, and sometimes binary, classification, in contrast to the multi-class classification used in this work.

The highest classification accuracy (91 %) was obtained for the model combining linear scaling of numeric data with frequency encoded categorical data. The calculated values of the Precision, Recall and F1-score metrics for this model have values 0.75, 0.778 and 0.757 respectively [36, 37]. These indicators can be considered acceptable, given the presence of a "*Data validation*" block at the output of the classifier. However, it is worth accepting the fact that the value of the Recall indicator, which is primary important in view of the need to provide the user with a greater number of possible ESMs, does not significantly exceed Precision and there is a sense in refining the model by increasing this indicator.

However, the question of the possibility and correctness of using methods of categorical data encoding that using dummy variables in solving the problem under study remains open. Possible reasons for such an outcome may be the fact that splitting one feature into several new ones entails introducing a large number of zeros and other constant values into the set. So, with the original number of features equal to 12, when using Binary Encoding, their number becomes equal to 59, and 99 when using Helmert or Backward-Difference Encoding, which also exceeds the total number of features in the original set equal to 34. Considering the small number of numeric features in the source dataset (22) and their normalization to the interval [0; 1], adding 89 new features (that is how many unique values all categorical features contain in general) consisting of zeros and ones makes the classification problem unsolvable.

CONCLUSION

As a result of the work, a dataset was formed from not only numeric, but also categorical data. The most suitable methods of data encoding, depending on their type, were determined experimentally: for numeric data – cleaning from outliers and scaling, for categorical data – frequency coding.

As a result of the experiments, the highest classification accuracy (91 %) was demonstrated by a combination of linear scaling of numeric data with frequency encoding of categorical data. Also, it demonstrates that introducing categorical data into the dataset, it became possible to achieve an actual increase in accuracy of 2 % compared to linear scaling without implementing categorical features, which can be considered a satisfactory result.

Acknowledgments

This work was supported by the "Priority 2023" Federal University Funding Program. The authors would also like to gratefully acknowledge High-Technology Center of BSTU named after V. G. Shukhov for provision of equipment for the research.

REFERENCES

1. Nekrasov S. A. (2023) Stimulating Electricity Consumption in Outsider Regions is a Necessary Condition for the Structural Stability of Russia. *Energetika. Izvestiya Vysshikh Uchebnykh Zavedenii i Energeticheskikh Ob'edinenii SNG = Energetika. Proceedings of CIS Higher Edu-*

cation Institutions and Power Engineering Associations, 66 (2), 186–200 (accessed 13 November 2024). https://doi.org/10.21122/1029-7448-2023-66-2-186-200 (in Russian).

2. Total energy consumption. *Enerdata*. Available at: https://energystats.enerdata.net/total-energy/world-consumption-statistics.html. (accessed 15 May 2024).
3. Electrical Balance and Electrical Energy Consumption in the Russian Federation From 2005 to 2022. *Federal State Statistics Service.* Available at: https://rosstat.gov.ru/storage/mediabank/elbalans_2022.xlsx. (accessed 15 May 2024) (in Russian).
4. Senshinova E. V., Zhuravlev A. E. (2021) Growth in Energy Consumption as One of the Challenges of the 21st Century. *Novaya nauka v novom mire: filosofskoe, sotsial'no-ekonomicheskoe, kul'turologicheskoe osmyslenie: Sbornik statei VIII Mezhdunarodnoi nauchno-prakticheskoi konferentsii, Petrozavodsk, 27 maya 2021 goda* [New science in the new world: philosophical, socio-economic, culturological understanding: Collection of articles of the VIII International scientific and practical conference, Petrozavodsk, May 27, 2021]. Petrozavodsk, 78–82 (in Russian).
5. Poletaev I. Yu., Androshina I. S. (2023) Formation of the Goals and Objectives of the New State Program for Energy Conservation and Energy Efficiency Improvement of the Russian Economy. *Herald of Russian Academy of Natural Sciences = Vestnik Rossiiskoi akademii estestvennykh nauk,* 23 (2), 120–124 (in Russian). https://doi.org/10.52531/1682-1696-2023-23-2-120-124.
6. Antonov A. N., Ryzhov G. A. (2010) Actual Issues of Increasing Efficiency and Investment Support for Energy Saving Programs at the Level of Municipalities in Russia. *Sovremennye tendentsii v ekonomike i upravlenii: novyi vzglyad. Sbornik dokladov IV Mezhdunarodnoi nauchno-prakticheskoi konferentsii. Ch. 1* [Modern trends in economics and management: a fresh approach. Proceedings of the IV International Scientific and Practical Conference. P. 1]. Novosibirsk, 59–66 (in Russian).
7. Ratner S. V. (2013) Issues of Practical Implementation of State Economic Policy in the Field of Energy Efficiency. *Economic analysis: theory and practice,* (29), 21–28 (in Russian).
8. *On Energy Saving and Improving Energy Efficiency, and on Amendments to Certain Legislative Acts of the Russian Federation*: Federal Law, 23.11.2009, No. 261, as amended in 13.06.2023. Available at: https://normativ.kontur.ru/document?moduleId=1&documentId=500643 (in Russian).
9. Nevokshenov A. Yu., Udovik A. V., Yurkovskaya G. I. (2015) Factors Influencing the Implementation of Energy Efficiency Programs and Energy Efficiency of Industrial Enterprises. *Sovremennye problemy ekonomicheskogo i sotsial'nogo razvitiya: mezhvuzovskii sbornik nauchnykh trudov. Vyp. 11* [Modern problems of economic and social development. Interuniversity collection of scientific papers. Iss. 11]. Krasnoyarsk, 32–34 (in Russian).
10. Lukishina L. V., Anisimov T. Yu., Mustafina O. N. (2017) Features of the Development of Energy Saving and Energy Efficiency Programs in the Context of Innovative Development of the Russian economy. *Innovatsionnoe razvitie rossiiskoi ekonomiki: materialy X Mezhdunarodnoi nauchno-prakticheskoi konferentsii. 25–27 oktyabrya 2017 g. T. 2* [Innovative development of the Russian economy: Proceedings of the X International scientific and practical conference: in five volumes, Moscow, October 25–27, 2017. Vol. 2]. Moscow, Plekhanov Russian University of Economics, 272–275 (in Russian).
11. Lebedeva N. A., Poletaeva L. P., Svezhintsev P. S. (2018) Adaptive Approach to the Formation of an Energy Saving Program. *Intellekt. Innovatsii. Investitsii = Intellect. Innovations. Investments,* (10), 24–27 (in Russian).
12. Li C. Z., Zhang L., Liang X., Xiao B., Tam V. W. Y., Lai X., Chen Z. (2022) Advances in the Research of Building Energy Saving. *Energy and Buildings*, 254, 111556. https://doi.org/10.1016/j.enbuild.2021.111556.
13. Mohsen M. S., Akash B. A. (2001) Some Prospects of Energy Savings in Buildings. *Energy Conversion and Management*, 42 (11), 1307–1315. https://doi.org/10.1016/s0196-8904(00)00140-0.
14. Popescu D., Bienert S., Schützenhofer C., Boazu R. (2012) Impact of Energy Efficiency Measures on the Economic Value of Buildings. *Applied Energy*, 89 (1), 454–463. https://doi.org/10.1016/j.apenergy.2011.08.015.
15. Song K., Ahn Y., Ahn J., Kwon N. (2019) Development of an Energy Saving Strategy Model for Retrofitting Existing Buildings: A Korean Case Study. *Energies*, 12 (9), 1626. https://doi.org/10.3390/en12091626.
16. Farkhutdinov R. R. (2017) Energy Saving Tools and their Application within the Framework of Regional Development Programs. *Alleya nauki = Alley of Science*, 3 (13), 555–559 (in Russian).
17. Koshlich Yu., Belousov A., Trubaev P., Grebenik A., Bukhanov D. (2020) Control Systems of Regional Energy Resources as a Digital Platform for Smart Cities. da Silva Bartolo P. J., da Silva F. M., Jaradat S., Bartolo H. (eds.). *Industry 4.0 – Shaping The Future of The Digital World*. London, CRC Press, 309–313. https://doi.org/10.1201/9780367823085-54.

18. Kuzmin A. M., Vysokovskaya E. A. (2019) The span Diagram is One of the Tools for Statistical Data Processing. *Metody menedzhmenta kachestva = Methods of Quality Management*, (11), 39 (in Russian).
19. Starovoytov V. V., Golub Yu. I. (2021) Data Normalization in Machine Learning. *Informatics,* 18 (3), 83–96 (in Russian). https://doi.org/10.37661/1816-0301-2021-18-3-83-96.
20. Ascione F., Bianco N., De Stasio C., Mauro G. M., Vanoli G. P. (2017) Artificial Neural Networks to Predict Energy Performance and Retrofit Scenarios for any Member of a Building category: A Novel Approach. *Energy*, 118, 999–1017. https://doi.org/10.1016/j.energy.2016.10.126.
21. Mokrousova E. S., Romodin A. V. (2010) The issue of Creation of a Mathematical Model of an Artificial Neural Network Within the Framework of the Development of an Automated System for Managing Energy Saving Programs. *Vestnik Permskogo gosudarstvennogo tekhnicheskogo universiteta. Elektrotekhnika, informatsionnye tekhnologii, sistemy upravleniya = Bulletin of the Perm State Technical University. Electrical engineering, information technology, control systems*, (4), 72–76 (in Russian).
22. Sednin A. V., Zherelo A. V. (2022) An Approach to Data Processing for the Smart District Heating System. *Energetika. Izvestiya Vysshikh Uchebnykh Zavedenii i Energeticheskikh Ob'edinenii SNG = Energetika. Proceedings of CIS Higher Education Institutions and Power Engineering Associations*, 65 (3), 240–249 (accessed 13 November 2024). https://doi.org/10. 21122/ 1029-7448-2022-65-3-240-249 (in Russian).
23. Zhang H., Feng H., Hewage K., Arashpour M. (2022) Artificial Neural Network for Predicting Building Energy Performance: A Surrogate Energy Retrofits Decision Support Framework. *Buildings*, 12 (6), 829. https://doi.org/10.3390/buildings12060829.
24. Albon C. (2018) *Machine Learning with Python Cookbook*. O'Reilly Media. 364.
25. Grus J. (2019) *Data Science from Scratch First Principles with Python*. O'Reilly Media. 403.
26. Categorical Variable Encoding. *Kaggle*. Available at: https://www.kaggle.com/code/harishvutukuri/categorical-variable-encoding (accessed 13 Juny 2024).
27. Processing of Categorical Features. *Habr*. Available at: https://habr.com/ru/articles/666234/ (accessed 13 Juny 2024) (in Russian).
28. Potdar K., Taher S., Chinmay D. (2017) A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. *International Journal of Computer Applications*, 175 (4), 7–9. https://doi.org/10.5120/ijca2017915495.
29. Novikova D. V. (2022) Comparative Analysis of the Effectiveness of Methods for Coding categorical Variables in the Problem of Predicting the Safety of Therapy for Multiple Sclerosis. *Zametki po informatike i matematike: Sbornik nauchnykh statei. Vyp. 14* [Notes on computer science and mathematics: Collection of scientific articles. Vol. 14]. Yaroslavl, Yaroslavl State University named after P.G. Demidov, 61–69 (in Russian).
30. Barkov D. V., Senotova S. A. (2021) Encoding of Categorical Features in Neural Networks. *Scientific Papers Collection of the Angarsk State Technical University*, 2021 (1), 3–8. https://doi.org/10.36629/2686-7788-2021-1-1-3-8 (in Russian).
31. Dong G., Liu H. (2018) *Feature Engineering for Machine Learning and Data Analytics*. CRC Press. 400. https://doi.org/10.1201/9781315181080.
32. Zheng A., Casari A. (2018) *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media. 215.
33. Ilyukovich-Strakovskaya A. M. (2015) Methods of Handle Categorial Features in Classification Problem. *Informatsionno-telekommunikatsionnye tekhnologii i matematicheskoe modelirovanie vysokotekhnologichnykh sistem: materialy Vserossiiskoi konferentsii s mezhdunarodnym uchastiem. Moskva, 20–24 aprelya 2015 g*. [Information and telecommunication technologies and mathematical modeling of high-tech systems: Proceedings of the All-Russian conference with international participation, Moscow, April 20–24, 2015]. Moscow, Peoples' Friendship University of Russia, 137–139 (in Russian).
34. Ezukwoke K. I. (2023) *Data Transformation for Machine Learning*. Available at: https://www.academia.edu/40436475/Data_Transformation_for_Machine_Learning.
35. *Keras API*. Available at: https://keras.io/ (accessed 22 Juny 2024).
36. *Scikit-learn*. Available at: https://scikit-learn.org (accessed 22 June 2024).
37. Sokolova M., Japkowicz N., Szpakowicz S. (2006) Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. Sattar A., Kang Bh. (eds). *AI 2006: Advances in Artificial Intelligence*. *Lecture Notes in Computer Science*, vol. 4304. Springer, Berlin, Heidelberg, 1015–1021. https://doi.org/10.1007/11941439_114.

Received: 16 January 2025 Accepted: 1 July 2025 Published online: 30 January 2026