# Research of algorithms of Data Mining

*Vitalii* Levkivskyi[1], *Nadiia* Lobanchykova[2,*], and *Dmytro* Marchuk[3]

[1]Zhytomyr Polytechnic State University, Department of computer science, 103 Chudnivska Str., Zhytomyr, 10005, Ukraine
[2]Zhytomyr Polytechnic State University, Department of computer engineering and cybersecurity, 103 Chudnivska Str., Zhytomyr, 10005, Ukraine
[3]Zhytomyr Polytechnic State University, Department of software engineering, 103 Chudnivska Str., Zhytomyr, 10005, Ukraine

**Abstract.** The article explores data mining algorithms, which based on rules and calculations, that allow us to create a model that analyzes the data provided by searching for specific patterns and trends. The purpose of this work is to analyze correlation-regression algorithms on a statistical dataset of chronic diseases. Data mining allows building many models, multiple algorithms can be used within a single solution. The article explores the algorithms of clustering, correlation analysis, Naive Bayes algorithm for obtaining different views of data. Since diabetes is one of the most dangerous chronic diseases, the pathogenesis of which is a lack of insulin in the human body, which causes metabolic disorders and pathological changes in various organs and tissues. As a result, it leads to disability of all functional systems of the body. It was decided to investigate the data related to this disease. Also, the quality of the developed methods of information retrieval from the dataset was evaluated and the most informative features were identified. The developed methods were implemented in the system of intellectual data processing. Past studies show promise of using data mining methods to improve the quality of patient care.

## 1 Introduction

The health care system is fully controlled by the state. According to the determined vectors and strategic vision of the Sustainable Development Strategy of Ukraine until 2030, one of the focuses is on providing an effective public health system, providing effective medical services, which is impossible without the use of modern information technologies [1]. The use of modern information technology is impossible without green IT [2].

The complexity of the experiments in this area is due to the object of the study, that is, human health. Most experiments on human health have legislative restrictions, as this can cause damage and prevent human rights. Therefore, the use of methods of data mining, in this case, for the analysis of already collected data can allow revealing a lot of hidden information, based on which further decisions can be made.

There are many algorithms for data mining. The purpose of this is to find patterns in the data. The knowledge gained through the methods of Data mining is accepted to be represented in the form of regularities (patterns) [3]. The article deals with classification algorithms and correlation-regression algorithms.

### 1.1 Analysis of published data

Volozhanin defines the basic concepts of correlation and calculates the correlation coefficient between the digital (numeric) value of each factor and the result of the sports. But the author does not use mathematical tools for analysis and does not provide specific research results [4]. The article [5] by the author team researched the detection of pathogenetic correlations between the parameters of standard laboratory methods of research in patients with iron deficiency anemia with chronic blood loss. Using the language R, a statistical analysis of the laboratory parameters was performed and the inherent correlations were established. Differences were found in correlation relationships between indicators, which reflect profound pathogenetic changes, using differentiation. These results have practical value for the assessment of the severity of metabolic damage and to monitor treatment effectiveness.

The systematic approach to the training of athletes is considered in the article [6]. Usage patterns of mathematical statistics methods in sports analysis are considered. Out of several methods, more reliable ones for detecting correlation relationships were chosen. The article uses a Fechner coefficient to determine the closeness of links between objects.

The article [7] presents an approach to the recognition of localized road signs using the method of support vector machine and histograms of oriented gradients. A method of plotting histograms of oriented gradients that do not depend on image size is proposed. For problems with multiple classes, the method of support vector machine is generalized. The experimental data show the advantages of the proposed method. The disadvantage of this study is that the experiment is not sufficiently covered.

The article of Andreeva [8] investigates the correlation analysis of the results of sociological researches to the

* Corresponding author: lobanchikovanadia@gmail.com

problem of attitude of the city population to the number of days off needed to celebrate the new year was explored. Statistical relationships were found between different feature groups. This study does not describe the algorithms that were used in the work.

### 1.2 Formulation of the problem

The purpose of the study is to use the algorithms of data mining for the processing of statistics and research into green IT creation and implementation processes:
- researching correlation-regression algorithms;
- the use of the support vector method;
- conducting research (exploratory) analysis;
- selection of the most unexpected dependencies.

## 2 Description of the data source

Unfortunately, medical statistics in Ukraine are not yet able to provide the necessary amount of statistical information for analysis and boast the extensive use of green IT. Therefore, the US Centers for Disease Control and Prevention (CDC) may be one of the sources for research. Chronic Disease Statistics in 500 U.S. cities were selected for data analysis based on information collected by the CDC [9]. This data is unique in itself because it covers 103 million people aged 18 and older and has 27,210,000 records across different statistical reporting territories, with populations ranging from 50 to 26,980. Also included in the data is the state, district, city, geographic coordinates, which further expands the analysis based on other statistics, for example, average household income, unemployment rate and more.

The indicators are divided into three main groups:

1. Unhealthy lifestyle (5 indicators).
2. Chronic diseases (13 indicators).
3. Population coverage by preventive methods (9 indicators).

All metrics are presented as a percentage of the population and a range of errors. The CDC, as a government organization, has sufficient resources to ensure that statistical information is collected and further processed. Chronic illness statistics are already a normalized database and contain information from verified sources.

**Naive Bayes classifier** belongs to a family (type) of probabilistic classifiers using the Bayes' theory. Despite its simplicity, the Naive Bayes classifier remains one of the most popular methods of solving the problem of text categorization, the problem of identifying documents that fall into one category or another. The purpose of the naive Bayesian classifier is to calculate the conditional probability by the formula (1):

$$p(C_k|x_1, x_2, \dots, x_n) \quad (1)$$

For each of $k$ the possible outcomes or classes $C_k$.

Let (may, if) $x = (x_1, x_2, \dots, x_n)$. Using Bayes' theorem, we can obtain, get the formula of the form (2):

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \propto p(C_k)p(x|C_k) = \\ = p(C_k, x_1, x_2, \dots, x_n) \quad (2)$$

The joint probability can be written as (3):

$$p(C_k|x_1, x_2, \dots, x_n) = \\ = p(x_1|x_2, \dots, x_n, C_k) \cdot p(x_2, \dots, x_n, C_k) = \\ = p(x_1|x_2, \dots, x_n, C_k) \cdot p(x_2|x_3, \dots, x_n, C_k) \cdot \\ \cdot p(x_3, \dots, x_n, C_k) = p(x_1|x_2, \dots, x_n, C_k) \cdot \\ \cdot p(x_2|x_3, \dots, x_n, C_k) \cdot p(x_n|C_k) \cdot C_k \quad (3)$$

Assume that all functions $x$ mutually independent, we can get the formula (4):

$$p(x_1|x_2, \dots, x_n, C_k) = p(x_1|C_k) \quad (4)$$

Therefore, the formula can be written as (5). So, it is the final formula for naive Bayesian classifiers. We offer training and testing in Python programming language.

$$p(C_k|x_1, x_2, \dots, x_n) \propto p(C_k, x_1, x_2, \dots, x_n) = \\ = p(x_1|C_k) \cdot p(x_2|C_k) \dots \cdot p(x_n|C_k) \cdot p(C_k) = \\ = p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \quad (5)$$

The training model is represented in the Python programming language is:

```
def train_dataset_for_column(column_to_predict):
  column_to_predict = column_to_predict
  train, test = train_test_split
    (data_only_crude_prev, test_size=0.2)
  best_chooser = BestSelectionTrainAlgorithm(
    lambda: linear_model.LinearRegression())
  best_chooser.max_best_indicators_count = 5
  best_chooser.low_score_for_best = 0.7
  best_chooser.max_combine = 2
  best_chooser.y_column_preparer =
    lambda data_frame: data_frame.astype(int)
  best_indicators =
best_chooser.best_indicators_from(train, test,
column_to_predict)
  descr = 'Best indicator for ' +
    column_to_predict + '\n'
  for train_result in best_indicators:
   train_result.columns_by_predict,accuracy=2.0)
    descr = descr + '
'.join(train_result.columns_by_predict) + ' ' +
column_to_predict + ': ' +
str(train_result.score) + '\n'
str(train_result.model.intercept_) + '\n'
  descr = descr + "\n\n"
  print(descr)
```

The testing model is represented in the Python programming language is:

```
def test_prediction_accuracy(model, test,
column_to_predict, column_by_predict,
accuracy=0.0):
  count_of_test_rows = len(test)
  count_of_close_predicted = 0
  test_by = test[column_by_predict].values
  test_to = test[column_to_predict].values
  predicted_values = model.predict(test_by)
  for i in range(0, count_of_test_rows):
    if isclose(predicted_values[i], test_to[i],
      abs_tol=accuracy/2):
      count_of_close_predicted += 1
  return count_of_close_predicted/
```

```
    count_of_test_rows
if __name__ == '__main__':
  mp.freeze_support()
  columns = data_only_crude_prev.columns.values
  p = mp.Pool(3)
  results = p.map(train_dataset_for_column,
columns)
```

Support vector machine provides significant accuracy with less computing power and can be used for both regression and classification problems. The purpose of the algorithm is to find a hyperplane in an *N*-dimensional space (*N* is the number of features) that clearly classifies data points (see Fig. 1).

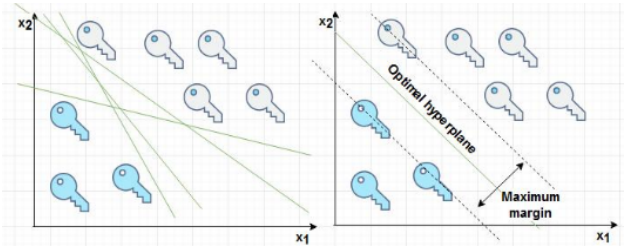


**Fig. 1**. Possible hyperplanes.

To separate the two classes of data points, there are many possible hyperplanes to choose from. The goal of the algorithm is to find the plane, which has a maximum margin (stock, reserve), that is, the maximum distance between data points of both classes. Maximizing the margin (stock, reserve) distance provides some reinforcement so that future data points can be classified with greater confidence.

Data points that fall on either side of the hyperplane can be classified into different classes. Also, the dimension of the hyperplane depends on the number of features (sighs, characteristics). If the number of input objects is 2, then the hyperplane is just a line. If the number of input objects is 3, then the hyperplane becomes a two-dimensional plane. It becomes difficult to imagine when the number of functions exceeds 3 [11].

The purpose of the algorithm is to maximize the difference between the data points and the hyperplane. A loss function that helps maximize margin is a hinged loss. The cost function and gradient renovation, (6):

$$c(x,y,f(x)) = \begin{cases} 0, & if.\, y * f(x) \geq 1 \\ 1 - y * f(x), & else \end{cases} \quad (6)$$

Cost is 0 if the predicted value and the actual value have the same sign. If this condition is not fulfilled – it is necessary to calculate the amount of damage. We also add a parameter of the regularization of the cost function. The task of the regularization parameter is to balance the maximization of margins and losses. After adding the regularization parameter, the cost functions are as follows, (7):

$$min_w\, \lambda\, ||w||^2 + \sum_{i=1}^{n} (1 - y_i(x_i, w))_+ \quad (7)$$

Once the loss function is known, it is possible to take private derivatives by weight to find gradients. Using gradients can be updated weight, (8), (9):

$$\frac{\delta}{\delta w_k} \lambda ||w||^2 = 2\lambda w_k \quad (8)$$

$$\frac{\delta}{\delta w_k}(1 - y_i\langle x_i, w\rangle)_+ = \begin{cases} 0, & if.\, y_i\langle x_i, w\rangle \geq 1 \\ -y_i x_{ik}, & else \end{cases} \quad (9)$$

When the model makes a mistake in the prediction of a data point class, it is necessary to include the loss along with the regularization parameter to update the gradient, (10):

$$\omega = \omega + a \cdot (y_i \cdot x_i - 2\lambda\omega) \quad (10)$$

When the model correctly predicts the data point class, all that is need to do is to update the gradient of the regularization parameter, (11):

$$\omega = \omega - a \cdot (2\lambda\omega) \quad (11)$$

The Best Selection Train Algorithm class was developed to implement the **support vector machine**.

For the selection of indicators that are significantly correlated with each other, this class has the following fields:

1. max_best_indicators_count – the maximum number of indicators to be selected.
2. low_score_for_best – a minimum estimate of the accuracy of the model (used to ensure that the model based on the model was considered the best).
3. max_combine – the number of elements in the array based on which the prediction is made (it should be noted that the long-term operation of the algorithm may be required when this indicator is of great importance).
4. y_column_preparer – the method used to prepare the test data (for example, the Bayesian algorithm cannot predict non-discrete values (such as float), so in this method, the Float value can be converted to Int).

The training model implements the following methods:

1. def fit(self, X, y) – a method for training the model, where **X** – an array of values that the training is based on, *y* is the expected value.
2. def score(self, X, y, sample_weight=None) – a method for estimating the quality of model prediction, where **X** – an array of values on which testing takes place, *y* is the expected value.

The model is built using the support-vector machine and depends only on the subset of the training data and in some cases gives the best result of regression analysis.

**Exploratory data analysis.** To investigate the relationship between the data, we use the Pearson correlation coefficient (Table 1), the values of which are interpreted based on absolute values. Possible values vary from 0 to ± 1. To evaluate the bond strength, a Cheddock table is usually used, according to which, absolute values less than 0.3 indicate weak bond strength, 0.3 to 0.5 – moderate, 0.5 to 0.7 - significant, 0, 7 – 0.9 - high, a value greater than 0.9 – very high. Figure 2 presents the matrix of correlation indicators (heat map) obtained by using linear regression and estimation of the results of the coefficient of determination (R-square).

After analyzing the indicators, some patterns have been identified that are logical. For example, a visit to a dentist leads to tooth loss. The lack of health insurance for

adults 18-64 years old (ACCESS2_CrudePrev) correlates with the prevalence of visits to a dentist or dental clinic for adults 18 years of age (DENTAL_CrudePrev). A correlation was found between such indicators as the prevalence of arthritis in adults 18 years of age (ARTHRITIS_CrudePrev) and high blood pressure in adults 18 years of age (BPHIGH_CrudePrev), a correlation coefficient of 0.75.

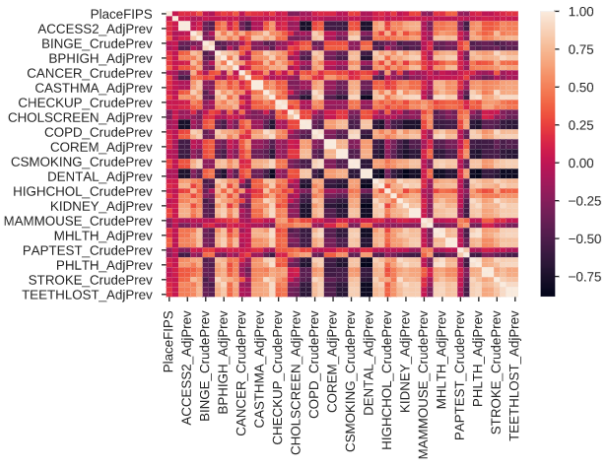


**Fig. 2.** Matrix of correlation indicators.

Correlations were also found between such indicators as the prevalence of arthritis among adults aged 18 and over (ARTHRITIS_CrudePrev) and chronic obstructive pulmonary disease in adults 18 years and older (COPD_CrudePrev) – correlation coefficient of 0,65.

Special attention has been paid to the correlation of diagnosed diabetes mellitus in adults aged 18 years (DIABETES) with other diseases. Diabetes mellitus is a disease that is included in the group of "leaders" along with cardiovascular and oncological diseases and is called the "disease of the century". If it was previously thought that diabetes is a disease of the elderly, then in the 21st century, it "became younger" and every year becomes "younger" and grows with a geometric progression.

This disease is in third place after cancer and cardiovascular. People are increasingly ill, even though medicine is improving day by day. The growing trend of cardiovascular, cancer, mental illness, and diabetes is observed not only in our country but also in more developed countries such as America. Table 1 presents some correlation indicators for the selected study object – DIABETES.

In the early stages of project development, it is often necessary to perform Exploratory data analysis (EDA) in order to identify patterns that identify the data. Data visualization helps to present large and complex datasets in a simple and visual way. Figure 3-4 shows the patterns between DIABETES_CrudePrev and other features.

The distribution density of two variables gives a gradient (Fig. 5), which in its direction indicates the direction of the greatest growth of a quantity whose value varies from one point of space to another and in magnitude is equal to the rate of growth of this value in this direction. The graphs presented in Figure 5 confirm the results of the studies. You can make sure that density of distribution of indicators DIABETES_CrudePrev and BPHIGH_CrudePrev, STROKE_CrudePrev, LPA_CrudePrev are better than DIABETES_CrudePrev and CASTHMA_CrudePrev.

**Table 1.** Correlation coefficients.

| | DIABETES_CrudePrev |
|---|---|
| ACCESS2_CrudePrev | 0.614854 |
| BPHIGH_CrudePrev | 0.846432 |
| BPMED_AdjPrev | 0.560538 |
| CASTHMA_CrudePrev | 0.357883 |
| ARTHRITIS_AdjPrev | 0.478496 |
| BPMED_CrudePrev | 0.572423 |
| CANCER_CrudePrev | 0.052244 |
| CHD_CrudePrev | 0.780102 |
| CHECKUP_CrudePrev | 0.514177 |
| CHOLSCREEN_CrudePrev | 0.100295 |
| COLON_SCREEN_CrudePrev | -0.640006 |
| COPD_CrudePrev | 0.696707 |
| CSMOKING_CrudePrev | 0.639347 |
| KIDNEY_CrudePrev | 0.900171 |
| LPA_CrudePrev | 0.832808 |
| OBESITY_CrudePrev | 0.729541 |
| SLEEP_CrudePrev | 0.716363 |
| STROKE_CrudePrev | 0.862107 |
| TEETHLOST_CrudePrev | 0.751177 |

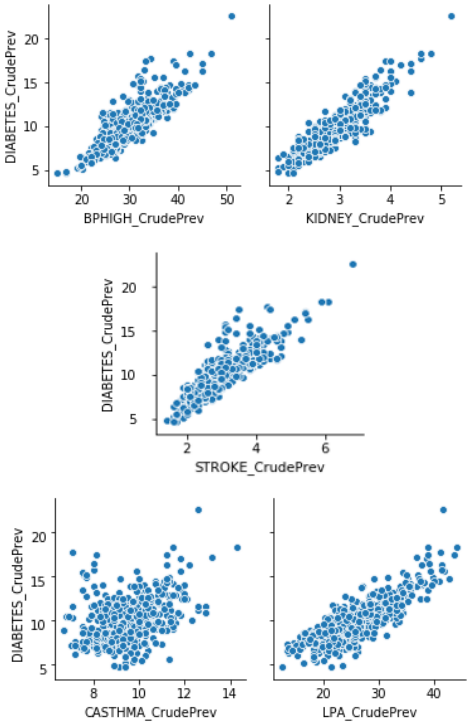


**Fig. 3.** Dependence between diabetes and other diseases.

To solve the problem of machine learning, the Scikit library was used. This library provides a wide variety of algorithms for (un)supervised learning.

The developed method [12] receives some elements on input, such as model, test data, name of the column for prediction, the names of the columns based on which the prediction is made, the desired accuracy of prediction (as absolute values), the way the test data is prepared.

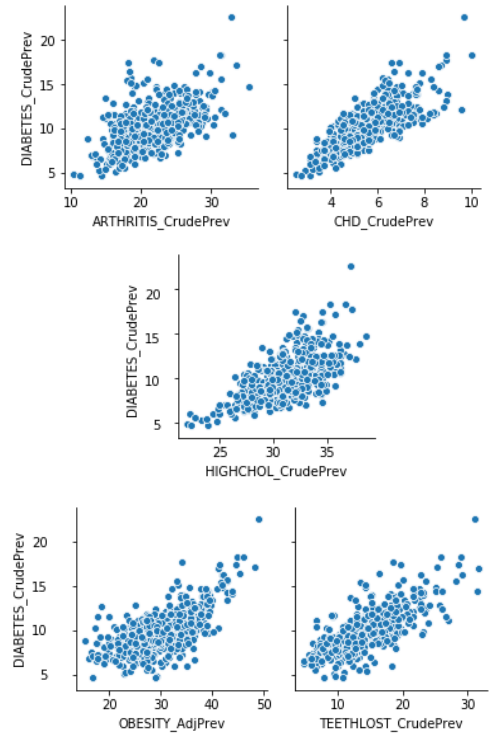


**Fig. 4**. Dependence between diabetes and other diseases.

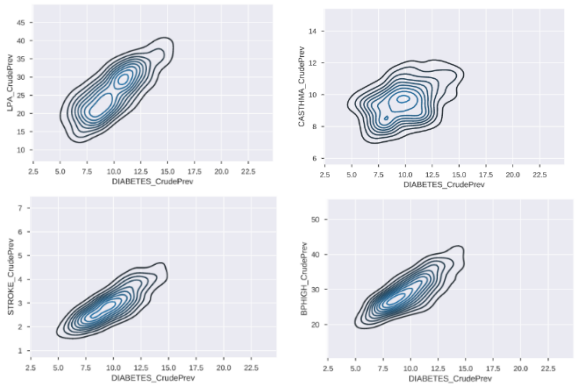

**Fig. 5**. Density distribution by parameters.

Also, the method of least squares for testing was used, which is defined as $1-\frac{u}{v}$, where u is the sum of squared errors of prediction:

$$\sum_{i=0}^{n}(y_{i\ \ actual}-\ y_{i\ \ predict})^2, \qquad (12)$$

and $v$ is the total sum of squares of the difference between the mean of the dependent variable and the exact value:

$$\sum_{i=0}^{n}(y_{i}\ \ -\ y_{i\ \ average})^2 \qquad (13)$$

As a result, a versatile method for estimating the quality of predicting was developed, that can be used to create green IT. Additionally, a deviation of ± 1%. As a result of the research, it is determined that various methods of data mining predict the prevalence of chronic diseases with sufficient accuracy.

## Conclusions

A correlation-regression analysis of chronic disease statistics was conducted based on the data set of the American Centers for Disease Control and Prevention (CDC). Models for Naive Bayes classifiers are developed. The support vector machine was used for classification and regression analysis.

An exploratory data analysis was performed, in which it was discovered unexpected dependencies between diagnosed diabetes mellitus among adults and high blood pressure, kidney disease, coronary artery disease and loss of all teeth.

As a result of the researches of the mentioned algorithms of data mining, models and methods were developed to establish the impact of some chronic diseases on others.

To increase the reliability of the results using the proposed models and methods, it is necessary to increase the amount of data. The studies conducted provide the basis for the practical design of energy efficient data processing centers.

## References

1. Stratehiia staloho rozvytku Ukrainy do 2030 roku. (Sustainable Development Strategy of Ukraine until 2030) (2020), https://www.undp.org/content/dam/ukraine/docs/SDGreports/UNDP_Strategy_v06-optimized.pdf. Accessed 15 Feb 2020
2. Green computing (2020) https://en.wikipedia.org/wiki/Green_computing. Accessed 15 Feb 2020.
3. B.K. Lebedev, V.B. Lebedev, O.B. Lebedev, Reshenie zadachi simvolnoy regressii metodami geneticheskogo poiska (The solution of the symbolic regression problem by genetic search methods). Izvestiya SFedU. Engineering Sciences, 212–224 (2015)
4. S.E. Volojanin, Opredelenie korrelyatsii mezhdu uprazhneniyami pauerliftinga i obschey fizicheskoy podgotovkoy (Determining of correlation between powerlifting exercises and general fitness). Vestnik Buriatskogo gosudarstvennogo universiteta 13, 39–43 (2011)
5. L.A. Pesotskaya, I.V. Yevstigneyev, T.V. Rublevskaya, A.A. Lukyanenko, V.S. Smirnov, Patogeneticheskie korrelyatsii laboratornyih pokazateley u bolnyih zhelezodefitsitnoy anemiey (Pathogenetic correlation of laboratory findings in patients with iron deficiency anemia). Actual Problems of the Modern Medicine: Bulletin of Ukrainian Medical Stomatological Academy **16**(4(56)), 171–175 (2016)
6. I.A. Osetrov, I.N. Nepryaev, Sravnitelnyie pokazateli korrelyatsii v sporte (Comparative correlation indicators in sports). Iaroslavskii pedagogicheskii vestnik 4–2009 (61), 60–64 (2009)

7. S.O. Lisitsyn, O.A. Baida, Raspoznavanie dorozhnyih znakov s pomoschyu metoda opornyih vektorov i gistogramm orientirovannyih gradientov (Recognition of road signs using the support vector method and oriented gradient histograms). Computer optics **36**(2), 289–295 (2012)
8. M.M. Andreeva, V.R. Volkov. Korrelyatsionnyiy analiz v sotsiologicheskih issledovaniyah (Correlation analysis in sociological). Herald of Kazan Technological University **7**, 271–274 (2013)
9. 500 Cities: City Boundaries (2019), https://chronicdata.cdc.gov/500-Cities/500-Cities-City-Boundaries/n44h-hy2j. Accessed 10 Feb 2020
10. Polynomial Regression (2018), https://towardsdatascience.com/polynomial-regression-bbe8b9d97491. Accessed 10 Feb 2020.
11. O.I. Sheremet, O.V. Sadovoi, Metod opornyih vektorovю (Support-vector machine (SVM)). Mathematical modeling 1(28), 13–17 (2013)
12. G.V. Marchuk, V.L. Levkivskyi, S.S. Kaliberda Intelektualniy analIz danih (Intelligent data analysis). Bionics of Intelligence **1**(92), 65–70 (2019)