УДК 519.237.7
DOI 10.34822/1999-7604-2020-4-33-41

АНАЛИЗ ВХОДНЫХ ПАРАМЕТРОВ ЭКСПЕРТНОЙ СИСТЕМЫ РАННЕЙ ДИАГНОСТИКИ ЗАБОЛЕВАНИЯ

А. С. Серобабов
Омский государственный технический университет, Омск, Россия
E-mail: aserobabow95@mail.ru

В статье рассмотрены вопросы подготовки входных параметров экспертной системы к этапу создания продукционных правил. Проведен анализ возможности понижения размерности (редуцирования) входных параметров с помощью метода главных компонент и факторного анализа. Оценка адекватности генеральной выборки для построения факторной модели проверена тестом Бартлетта и критерием Кайзера – Мейера – Олкина. Модель подвергнута качественным оценкам (доля объясненной дисперсии, общность факторов и интерпретируемость модели). По полученным факторам построены графики корреляционной связи с входными параметрами и представлены в виде карты взаимосвязей. В результате получена модель, состоящая из двух факторов, которая свидетельствует, что входные параметры системы имеют сложную взаимосвязь и не могут быть редуцированы.

Ключевые слова: факторный анализ, метод главных компонент, тест Бартлетта, критерий Кайзера – Мейера – Олкина, экспертная система диагностики заболевания.

ANALYSIS OF INPUT PARAMETERS OF THE EXPERT SYSTEM FOR EARLY DIAGNOSIS OF THE DISEASE

A. S. Serobabov
Omsk State Technical University, Omsk, Russia
E-mail: aserobabow95@mail.ru

The article describes the preparation of input parameters of the expert system for the construction phase of production rules. The analysis of the possibility of dimensional reduction of input parameters by factor analysis methods such as principal component analysis and factor analysis is made. The adequacy assessment of the general sample for constructing a factor model is verified by Bartlett's test and the Kaiser-Meyer-Olkin Test. The model is subjected to qualitative assessments: the proportion of explained variance, the generality of factors, and the interpretability of the model. Based on the obtained factors, correlation graphs with input parameters are plotted, and graphically represented as a map of relationships. As a result, a model consisting of two factors is obtained, which indicates that the input parameters of the system have a complex correlation and cannot be reduced.

Keywords: factor analysis, principal component analysis, Bartlett's test, Kaiser-Meyer-Olkin Test, expert system for diagnosis of disease.

Введение

Широкое применение экспертных систем в автоматизации задач, связанных с обработкой, прогнозированием, обучением на основе биомедицинской информации, является приоритетным направлением, которое, несмотря на обилие разработанных медицинских экспертных систем диагностики заболеваний печени [1–5], развивается благодаря появлению новых алгоритмов проектирования систем, получению качественно новых синергетических связей со смежными областями искусственного интеллекта.

Наибольшую сложность в построении экспертной системы представляет этап выявления и создания экспертных правил. Правила могут обладать большим количеством уникальных условий, что связано с количеством рассматриваемых входных переменных, специфи-

кой исследования и неправильным подходом к работе с данными. Чтобы избежать построения сложной базы правил там, где можно обойтись меньшим набором условий, применяются математические методы. Они позволяют сконцентрировать исходную информацию и выражают большее число рассматриваемых факторов через меньшее число более емких, содержащих большое количество информации, внутренних факторов, которые, однако, не поддаются непосредственному измерению. В результате будут получены новые знания о взаимосвязях внутри генеральной выборки, повысится эффективность процесса постановки диагноза, что, в свою очередь, повлияет на качество и продолжительность жизни пациентов.

Для этой цели выбран метод машинного обучения без учителя – факторный анализ. Факторный анализ позволяет сконцентрировать исходную информацию, используя меньшее количество признаков, которые включают в себя характеристики других признаков. Применимость факторного анализа в данном исследовании обусловлена распространенностью его использования в других работах [6–8]. С помощью него можно выявить внутренние факторы, отвечающие за наличие линейных статистических корреляций между исследуемыми параметрами.

Предлагается подход с использованием факторного анализа как инструмента исследования ранней диагностики заболевания, нацеленный на определение взаимосвязи между данными, выявление латентных факторов, редуцирование количества переменных, охватывающих дисперсию входных параметров [9].

Материал и методы

В качестве исходных данных использованы результаты, которые получены в результате обследования 149 пациентов с неалкогольной жировой болезнью печени. Все пациенты, участвующие в данном исследовании, отобраны в результате диспансеризации населения из различных поликлинических учреждений города Омска и переданы для исследования связей между лабораторными параметрами и стадией заболевания.

Результаты подвергнуты первичной обработке и вычислению корреляционных связей между стадией заболевания и лабораторными результатами [10]. По результатам предыдущих исследований [10–12] выделены три основных параметра, которые выступают входными данными экспертной системы: L_{lep} – лептин, L_{ObR} – рецептор лептина, D_{NASH} – наличие неалкогольного стеатогепатита.

Все вычисления выполнялись на языке Python с использованием среды разработки Jupyter Notebook, включающей библиотеки машинного обучения и анализа данных, а также программные инструменты создания рисунков для наглядного представления выявленных функциональных и корреляционных зависимостей.

Первоначальная выборка данных из исследования [10] имеет пропуски значений, в связи с этим необходимо исключить кортежи, у которых присутствуют пропуски. Для этого используется метод библиотеки pandos.dropna, который преобразует исходный набор данных из 149 упорядоченных кортежей в набор данных без пропусков («очистка данных»), состоящий из 64 оставшихся кортежей.

Фрагмент результата выполнения операции «очистки данных» из результатов, полученных в исследовании [10], представлен в табл. 1. В таблице отражены три входных параметра, где все значения – положительные и не имеют пропусков и выбросов. Следующий шаг – понижение размерности входной выборки методами факторного анализа.

Таблица 1

Фрагмент данных после исключения пациентов с пропусками значений

№	Значение L_{lep} нг/мл	Значение L_{ObR} нг/мл	Стадия D_{NASH}
1	30,467	4,485	2
2	13,553	10,893	1
3	3,567	12,366	2
4	16,855	4,125	1
5	10,402	4,374	1

Окончание табл. 1

№	Значение L_{lep} нг/мл	Значение L_{ObR} нг/мл	Стадия D_{NASH}
…	…	…	…
57	33,018	5,643	1
61	108,8	3,852	2
62	5,349	2,46	1
63	56,161	7,26	2
64	30,633	7,686	2

Примечание: составлено автором на основании данных, полученных в исследовании.

Постановка задачи

В рамках данной работы будет рассмотрено применение факторного анализа для редукции признакового пространства входных параметров экспертной системы.

Из терминов факторного анализа пространство признаков представляется многомерным (k-мерным, где k – количество исходных признаков, вошедших в исследование). Каждый объект (пациент) представляется точкой в этом пространстве и имеет вид:

$$X_t = \{L_{\text{lep } t}, L_{\text{ObR } t}, D_{NASH\ t}\}, \quad (1)$$

где t = 1, 2, …, N, N – число исследуемых объектов (пациентов).

Для решения задачи редуцирования входных переменных необходимо предположить, что X_i, $i = \overline{1,k}$ линейно зависят от F, где $F = (F_1, F_2)$ – множество предполагаемых факторов (восприимчивость к лептину и наличие воспалительных процессов в печени), m= {1,2}.

$$X = \Lambda F, \quad (2)$$

где $\Lambda = \begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1m} \\ \vdots & \ddots & \vdots \\ \lambda_{k1} & \cdots & \lambda_{km} \end{pmatrix}$ – множество факторных нагрузок для полученной модели.

Полученные факторы должны быть ортодоксальны друг другу и вбирать в себя наибольшее значение дисперсии.

В рамках данного исследования использованы два метода факторного анализа:

1) метод главных компонент (далее – МГК), в котором наблюдаемые значения каждого из признаков X_i , $i = \overline{1,k}$ представляются в виде линейных комбинаций факторных нагрузок λ_{ij} и факторов F_j, где j = 1, 2 … m, m – количество факторов:

$$X_i = \sum_{j=1}^{m} a_{ij} F_j; \quad (3)$$

2) модель собственного факторного анализа (далее – ФА), при которой наблюдаемые значения определяются не только факторами, но и действием локальных случайных причин:

$$X_i = \sum_{j=1}^{m} a_{ij} F_j + u_j. \quad (4)$$

Оценка возможностей использования анализа главных компонент и факторного анализа

С точки зрения формальной оценки качества факторной модели используются критерий адекватности Кайзера – Мейера – Олкина (далее – КМО) и тест Бартлетта.

Критерий КМО применяется для оценки применимости факторного анализа к данной выборке и вычисляется по формуле 5:

$$KMO = \frac{\sum_{i=1}^{p}\sum_{\substack{j=1 \\ i\neq j}}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{\substack{j=1 \\ i\neq j}}^{p} p_{ij}^2 + \sum_{i=1}^{p}\sum_{\substack{j=1 \\ i\neq j}}^{p} r_{ij}^2}, \quad (5)$$

где $p_{ij} = \frac{R_{ij}}{\sqrt{R_{ii} \cdot R_{jj}}}$ – коэффициент парциальной корреляции;

$r_{ij} = R(X_i, X_j)$ – корреляция Пирсона.

Для вычисления критерия КМО используется библиотека FactorAnalyzer [13] и метод calculate_kmo. На основании данных, полученных на этапе «очистки» из табл. 1, вычислен критерий КМО. Результат и фрагмент программы исследования представлены на рис. 1.

```
In [170]: from factor_analyzer import FactorAnalyzer

In [171]: from factor_analyzer.factor_analyzer import calculate_kmo

In [176]: kmo_all,kmo_model=calculate_kmo(df_scaled)

In [184]: kmo_model
Out[184]: 0.6476798446824265
```

Рис. 1. Фрагмент листинга программы проверки генеральной выборки соответствия к критерию адекватности КМО
Примечание: скриншот автора.

Полученное значение КМО = 0,65 можно интерпретировать с помощью табл. 2 из работы Кайзера – Мейера – Олкина [14]. Первый столбец представляет собой диапазоны численного значения критерия КМО, второй столбец – словесная интерпретация значения. Результаты теста сопоставлены с диапазоном значений и интерпретированы как «сомнительные» или «приемлемые». Следовательно, данные применимы для построения, но с некоторой оговоркой.

Таблица 2

Интерпретация меры выборочной адекватности Кайзера – Мейера – Олкина [14]

Диапазоны значений	Степень применимости факторного анализа
От 0,9 до 1	Отличная
От 0,8 до 0,9	Хорошая
От 0,7 до 0,8	Приемлемая
От 0,6 до 0,7	Сомнительная
От 0,5 до 0,6	Малопригодная
От 0 до 0,5	Недопустимая

Тест Бартлетта проверяет возможность построения факторной модели и используется для проверки предположения, что выборки имеют равные дисперсии. Для этого производятся вычисления по формулам 6–8 [15]:

$$T = \frac{M}{c}. \tag{6}$$

$$M = (N-k)\cdot \ln\left(s_p^2\right) - \sum_{i=1}^{k}(n_i - 1)\cdot \ln\left(s_p^2\right). \tag{7}$$

$$c = 1 + \frac{1}{3\cdot(k-1)}\cdot\left(\sum_{i=1}^{k}\left(\frac{1}{n_i-1}\right) - \frac{1}{(n-k)}\right), \tag{8}$$

где k – количество выборок;

n_i – объем выборки ($i = \overline{1,k}$);

$N = \sum_{i=1}^{k} n_i$;

$s_p^2 = \frac{1}{N-k}\cdot\sum_i^k n_i - 1 \cdot s_i^2$ – суммарная оценка дисперсии;

$s_i^2 = \frac{1}{n_i-1}\cdot\sum_{j=1}^{n_i}(X_{ij} - \overline{X_\iota})^2$;

$\overline{X_i} = \frac{1}{n_i} \cdot \sum_{j=1}^{n_i} X_{ji}$.

Для программного вычисления теста Бартлетта используется метод calculate_bartlett_sphericity, реализованный в библиотеке FactorAnalyzer [13]. На основании данных, полученных на этапе «очистки» из табл. 1, вычислен результат теста Батлетта. Результат и фрагмент листинга программы изображены на рис. 2.

```
In [174]: from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity
          chi_square_value,p_value=calculate_bartlett_sphericity(df_scaled)
          chi_square_value, p_value

Out[174]: (6.335176822487419, 0.0422788599867906)
```

Рис. 2. Фрагмент листинга программы с результатами теста Бартлетта
Примечание: скриншот автора.

Первое вычисленное значение является критерием хи-квадрата x^2 = 6,335. По полученному значению также вычисляется значение p-value. Так как полученное значение p-value < 0,05, то принимается гипотеза H_1 – корреляционная матрица не диагональная, следовательно, можно строить факторную модель.

Результаты. По результатам тестов генеральная выборка удовлетворяет критериям для проведения факторного анализа с построением модели, приведенной к новым факторам: восприимчивость к L_{lep} и наличие воспалительных процессов в печени.

Первый шаг в факторном анализе – центрирование и нормирование исходных значений признаков выборочной совокупности с помощью преобразования:

$$X_{jt}^{\text{Ц}} = \frac{X_{jt}^{\text{исх}} - \overline{X_J}}{\sigma_j}, \quad (9)$$

где $X_{jt}^{\text{исх}}$ – исходное значение j-ого признака;

$\overline{X_J}$ – среднее значение j-ого признака;

σ_j – стандартное отклонение j-ого признака.

Далее для вычисления факторных нагрузок и доли объясненной дисперсии методом главных компонент используется класс PCA библиотеки sklearn [16]. Входными данными являются значения, полученные в табл. 1. Результаты новой модели изображены на рис. 3a, 3b. Так, на рис. 3a представлен график собственных значений полученной модели факторов F_1, F_2. Для оценки пригодности полученных факторов используется критерий «каменистой осыпи». Его основная цель заключается в выявлении такой точки, в которой убывание собственных значений замедляется наиболее сильно. Это дает основание предполагать, что дальнейшее включение искусственных переменных качественно не улучшает построенную модель. Так как на графике изображено всего две точки и их отличие невелико, то, исходя из определения критерия, оба фактора принимаются как значимые. Далее применим критерий Кайзера, который устанавливает границу выбора факторов с собственными значениями больше единицы. На рис. 3a построена прямая, равная $y = 1$. Исходя из этого, при пересечении прямой собственных значений с прямой, характеризующей эмпирический критерий Кайзера, фактор должен быть исключен из построенной модели. Отсюда следует, что факторов с собственным значением меньше единицы нет, значит, оба фактора остаются. Дальнейшая интерпретация результатов основывается на критерии объясненной дисперсии. Доля объясненной дисперсии вычисляется по формуле:

$$D_{\text{об}} = \sum_{i=1}^{k} \frac{\gamma_i}{k}, \quad (10)$$

где k – количество переменных;

γ_i – i -е собственное число.

По формуле 10 получен результат $D_{\text{об}}$ = 0,777. Для наглядного представления построен кумулятивный график доли объясненной дисперсии, изображенный на рис. 3b.

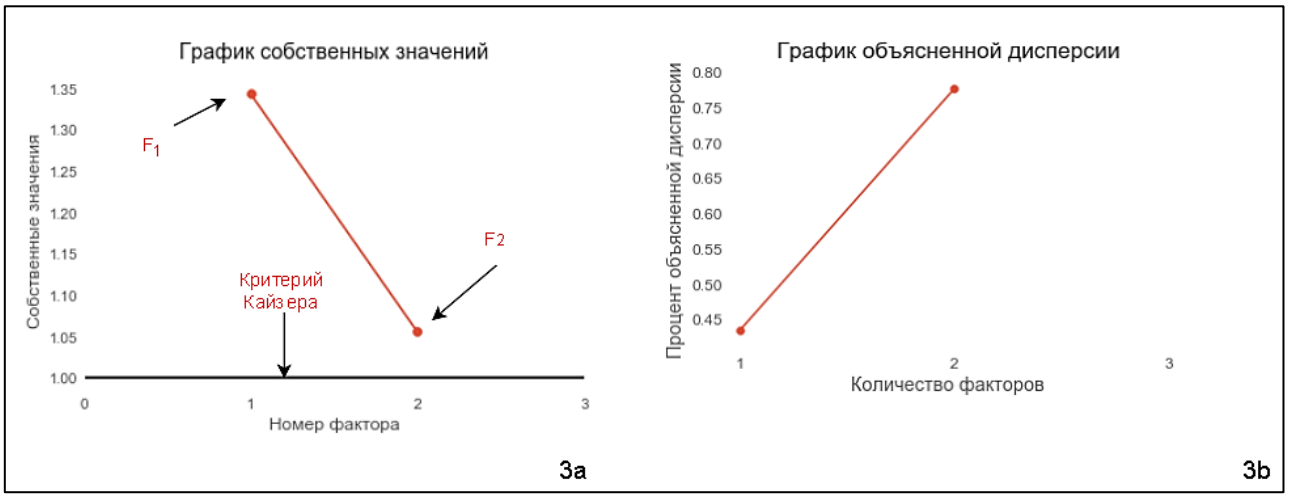

Рис. 3. Результаты собственных значений и доля объясненной дисперсии по полученной модели:
a – график собственных значений для факторов F_1, F_2;
b – кумулятивный график накопленной объясненной дисперсии
Примечание: составлено автором.

Полученное значение $D_{\text{об}}$ больше 0,5, значит, доля объясненной дисперсии больше, чем полученный остаток необъясненности. Отсюда следует, что полученную модель можно использовать с учетом того, что 20 % дисперсии параметров $L_{\text{lep }t}, L_{\text{ObR }t}, D_{NASH\ t}$ будет утеряно при переходе из трехмерного пространства $X_i = \{L_{\text{lep }t}, L_{\text{ObR }t}, D_{NASH\ t}\}$ к двумерному пространству признаков $X_i = \{F_1, F_2\}$, где X_i – i-й пациент генеральной выборки.

Для оценки интерпретируемости модели, построенной по методу главных компонент, строится корреляционная матрица входных факторов с исследуемыми параметрами. На рис. 4 представлен график корреляционной матрицы A размерности $m \times n$, где m – количество новых факторов, полученных из модели; n – количество параметров, используемых для постановки диагноза пациента. Источником данных являются значения модели, полученные в результате построения модели методом главных компонент. На пересечении столбца и строки выбирается значение, которое является значением корреляции между двумя выбранными переменными. В множестве A полученные факторы невозможно однозначно истолковать относительно входных параметров. Так, F_1 имеет схожую корреляцию с L_{lep} и D_{NASH}, отличающуюся лишь знаком, но не силой связи между ними. На основании шкалы Чеддока [17] связь параметров L_{lep} и D_{NASH} с F_1 интерпретируется как «слабосвязанная», с L_{obr} – как «заметная». F_1 имеет корреляцию со всеми тремя параметрами, поэтому он не имеет словесной интерпретации. Фактор F_2 не несет в себе определенного смысла, кроме связи воспалительного процесса с количественным содержанием L_{lep} у пациента. Отсюда следует вывод, что построенная модель не интерпретируема, входная выборка не может быть редуцирована.

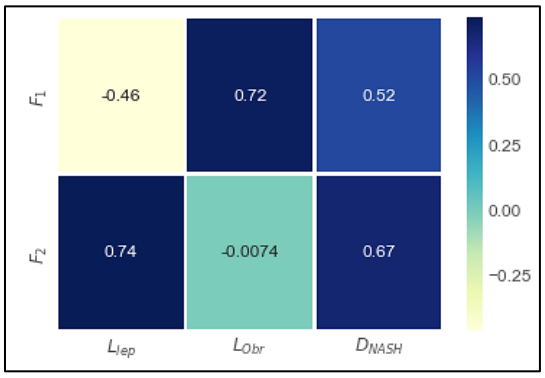

Рис. 4. Карта взаимосвязей с новыми факторами в методе главных компонент
Примечание: составлено автором.

Для вычисления факторных нагрузок и необъясненной уникальности факторов методом факторного анализа построенной модели используется класс FactorAnalysis библиотеки sklearn [13]. Результат необъясненной уникальности факторов представлен на рис. 5, на котором видно, что входные переменные объяснены плохо. Причина этого – высокое значение уникальности исследуемых параметров.

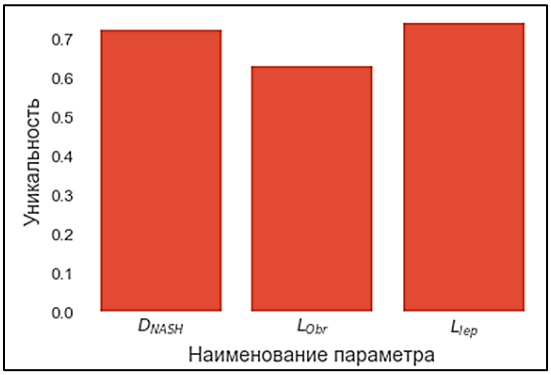

Рис. 5. Диаграмма необъясненной уникальности параметров после построения факторного анализа
Примечание: составлено автором.

Для оценки интерпретируемости модели, построенной по методу факторного анализа по данным из табл. 1, строится карта признаков, изображенная на рис. 6. Так, фактор F_1 имеет корреляцию 0,61 с признаком Obr, качественно характеризующую связь по шкале Чеддока как «заметную». Остальные связи являются «умеренными». Фактор F_2 и вовсе не имеет заметных связей. Это означает, что данный фактор не объясняет входные данные. Отсюда следует вывод, что построенная модель неинтерпретируема, входная выборка не может быть редуцирована.

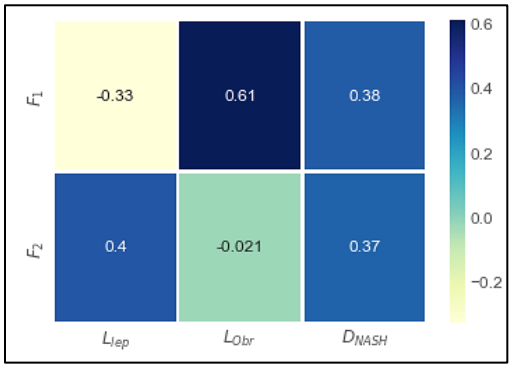

Рис. 6. Карта взаимосвязей с новыми факторами в методе факторного анализа
Примечание: составлено автором.

Обсуждение и заключение

Проведено исследование на наличие взаимосвязей входных параметров экспертной системы ранней диагностики заболевания, нацеленное на понижение размерности данных на вход экспертной системы.

Генеральная совокупность проверена тестом Бартлетта, и получено значение p-value < 0,05. Это означает, что корреляционная матрица является неотрицательной, и данные могут использоваться для проведения факторного анализа. Также входная выборка удовлетворяет критерию Кайзера – Мейера – Олкина. По этой причине данные признаются адекватными.

В результате применения методов факторного анализа построены две модели: одна – на основе метода главных компонент, вторая – на основе факторного анализа. В каждой модели получены факторы: F_1 – восприимчивость к лептину и F_2 – наличие воспалительных процессов в печени.

На основе полученных количественных результатов, метода главных компонент и факторного анализа построены карты взаимосвязей, которые, в свою очередь, являются графической формой представления матрицы корреляции с искусственно введенными факторами. В ходе оценки полученных моделей сделаны выводы, что обе модели неинтерпретируемы и входная выборка не может быть редуцирована. Следовательно, нечеткие процедурные правила для базы знаний должны браться из эмпирического опыта экспертов прикладной области.

Полученные результаты будут использоваться при работе экспертной системы ранней диагностики заболевания в качестве ограничений, налагаемых на составление продукционных правил системы в связи с невозможностью редуцирования входных параметров (L_{lep}, L_{ObR}, D_{NASH}) и наличием сложных нелинейных взаимосвязей.

Перспективным представляется применение в других исследованиях, где в качестве объекта исследования выступает один из перечисленных лабораторных параметров.

Литература

1. Polat K., Şahan S., Kodaz H., Güneş S. Breast Cancer and Liver Disorders Classification using Artificial Immune Recognition System (AIRS) with Performance Evaluation by Fuzzy Resource Allocation Mechanism // Expert Systems with Applications. 2007. Vol. 32. Iss. 1. P. 172–183.

2. Rizzo L., Longo L. An Empirical Evaluation of the Inferential Capacity of Defeasible Argumentation, Non-Monotonic Fuzzy Reasoning and Expert Systems // Expert Systems with Applications. 2020. Vol. 147. P. 113220. DOI 10.1016/j.eswa.2020.113220.

3. Osuagwu C. C., Okafor E. Framework for Eliciting Knowledge for a Medical Laboratory Diagnostic Expert System // Expert Systems with Applications. 2010. Vol. 37, Iss. 7. P. 5009–5016. DOI 10.1016/j.eswa.2009.12.012.

4. Abdar M., Zomorodi-Moghadam M., Das R., Ting I-H. Performance Analysis of Classification Algorithms on Early Detection of Liver Disease // Expert Systems with Applications. 2017 Vol. 67. P. 239–251.

5. Altay E. V., Alatas B. A Novel Clinical Decision Support System for Liver Fibrosis using Evolutionary Multi-Objective Method Based Numerical Association Analysis // Medical Hypotheses. 2020. Vol. 144. P. 110028.

6. Weeraratne N. C. The Effectiveness of Factor Analysis as a Statistical Tool of Variable Reduction Technique // International Journal of Core Engineering and Management. 2016. Vol. 3. P. 145–153.

7. Вирстюк А. Ю., Микшина В. С. Применение факторного анализа для редукции признакового пространства нагнетательных скважин // Вестник кибернетики. 2018. № 2. С. 172–178.

8. Li S., Sari Y. A., Kumral M. New Approaches to Cognitive Work Analysis through Latent Variable Modeling in Mining Operations // International Journal of Mining Science and Technology. 2019. Vol. 29. Iss. 4. P. 549–556. DOI 10.1016/j.ijmst.2019.06.014.

9. Баранов В. В., Белоновская И. Д., Чепасов В. И. Факторный анализ как инструмент педагогического знания о саморазвитии студента университетского комплекса // Вестн. Оренбург. гос. ун-та. 2012. № 2. С. 145–148.

10. Серобабов А. С., Чебаненко Е. В., Денисова Л. А., Кролевец Т. С. Разработка экспертной системы ранней диагностики заболеваний: программные средства первичной обработки и выявление зависимостей // Омск. науч. вестн. 2018. № 4 (160). С. 179–184.

11. Серобабов А. С. Формирование диапазонов переменных экспертной системы с использованием дерева принятия решений // Journal of advanced research in technical science. 2019. № 17–2. С. 161–166.

12. Серобабов А. С. Проверка входных параметров экспертной системы на соответствие нормальному закону распределения // Проблемы и перспективы студен. науки. 2019. № 2 (6). С. 3–6.

13. Factor Analyzer package. URL: https://factor-analyzer.readthedocs.io/en/latest/factor_analyzer.html (дата обращения: 11.10.2020).

14. Kaiser H. F. An Index of Factorial Simplicity // Psychometrika. 1974. Vol. 39. No. 1. P. 31–36.

15. Barlett M. S. Properties of Sufficiency and Statistical Tests // Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences. 1937. Vol. 160. P. 268–282.

16. Principal component analysis. URL: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html (дата обращения: 11.10.2020).

17. Костюченко О. А. Анализ математической модели объёма производства продукции и прогнозирование выручки // Концепт. 2014. № 3. URL: https://cyberleninka.ru/article/n/analiz-matematicheskoy-modeli-obyoma-proizvodstva-produktsii-i-prognozirovanie-vyruchki (дата обращения: 12.10.2020).