ISSN 2713-1734

Научная статья
УДК 51-74
EDN VDHPRA
DOI 10.17150/2713-1734.2023.5(2).172-191

П.С. Щербань
Балтийский федеральный университет им. И. Канта,
г. Калининград, Российская Федерация

Р.В. Абу-Хамди
Балтийский федеральный университет им. И. Канта,
г. Калининград, Российская Федерация

А.Н. Карагадян
Балтийский федеральный университет им. И. Канта,
г. Калининград, Российская Федерация

Использование библиотеки Scikit-learn и собственной программы по кластеризации для обработки статистических данных по отказам двухконтурных газовых котлов

Аннотация. На современном этапе развития техники у большого числа промышленных предприятий и сервисных организаций часто возникает проблема обработки массивов статистических данных по аварийным ситуациям и отказам оборудования. Рост числа информации и ее сложность, многофакторность возникающих отказов и взаимосвязанные с этим риски — не позволяют обходиться примитивными механизмами статистического анализа и учета. Одним из способов стратифицировать поступающие данные, выявить определенные закономерности в отказах оборудования и установить весомость влияющих факторов является кластеризация статистической информации. Для малых и средних предприятий, задействованных в процессе технического обслуживания и ремонта это особенно важно, поскольку наряду с выявлением причин возникновения неисправностей и отказов, применение кластеризации к массиву статистических данных дает информацию для стратегического планирования — объема ремонтных работ, объема закупки материалов и комплектующих, управления временем и в целом прогностики. Обработка массива данных требует привлечения соответствующего программного обеспечения. Для этого могут быть использованы как имеющиеся в свободном доступе пакеты (например, использование в работе библиотеки scikit-learn на интерактивной вычислительной среде Jupyter Notebook), так и собственные программные продукты, сформированные для оценки конкретной проблемы. В исследовании проводиться последовательная обработка и нормализация исходных данных по отказам котелкового оборудования для проведения кластерного анализа, далее выполняется сам кластерный анализ и сравниваются его результаты, полученные на базе общедоступной библиотеки scikit-learn на интерактивной вычислительной среде Jupyter Notebook и программы по кластеризации данных собственной разработки.

Ключевые слова. Математические и аналитические виды анализа, кластерный анализ, программная обработка данных, Python, управление качеством, отказы газовых систем.

Информация о статье. Дата поступления: 1 февраля 2023 г.; дата принятия к публикации: 14 апреля 2023 г.; дата онлайн-размещения: 14 мая 2023 г.

Original article

P.S. Shcherban
Immanuel Kant Baltic Federal University, Kaliningrad, Russian Fedaration
R.V. Abu-Hamdi
Immanuel Kant Baltic Federal University, Kaliningrad, Russian Fedaration
A.N. Karagadian
Immanuel Kant Baltic Federal University, Kaliningrad, Russian Fedaration

Using the Scikit-Learn Library And Our Own Clustering Program to Process Statistical Data on Failures of Double-Circuit Gas Boilers

Abstract. At the present stage of technology development, a large number of industrial enterprises and service organizations often have the problem of processing arrays of statistical data on emergency situations and equipment failures. The increase of information quantity and its complexity, the multifactorial nature of the occurred failures and the risks, which are associated with it, do not allow us to use primitive mechanisms of statistical analysis and accounting. Clustering of statistical information is one of the ways to stratify incoming data, identify certain patterns in equipment failures and establish the weight of influencing factors. For small and medium–sized enterprises involved in the maintenance and repair process, this is especially important, because along with identifying the causes of malfunctions and failures, the application of clustering to an array of statistical data provides information for strategic planning — the volume of repair work, the volume of materials and components procurement, time management, and in general, prognostics. Processing of an array of data requires the involvement of appropriate software. For this purpose, both freely available packages (for example, the scikit-learn library with the Jupyter Notebook interactive computing environment used in the work) and proprietary software products designed for investigation of specific problem can be used. In the study, we conducted sequential processing and normalization of the initial data on boilers equipment failures for cluster analysis. Then the cluster analysis itself is performed and its results are compared, obtained on the basis of the scikit-learn library with Jupyter Notebook public interactive computing environment and the data received from clustering program of our own design.

Keywords. Mathematical and analytical analysis, cluster analysis, software data processing, Python, quality management, gas system failures.

Article info. Received 1 February , 2023; Accepted 14 April, 2023; Available online 14 May, 2023.

Введение

Несмотря на значительный рост качества производства энергетического и нефтегазового оборудования, появление новых материалов и технологических решений, проблема пост продажного технического обслуживания и контроля за техническим состоянием остается по-прежнему крайне актуальной. Крупные предприятия топливно-энергетического комплекса часто решают данные

вопросы либо создавая собственные структурные подразделения по ремонту и техническому обслуживанию (включающие в себя как правило научно-исследовательский или аналитический отдел и лаборатории), либо заключая договора сопровождения со сторонними предприятиями, деятельность которых напрямую связана со сбором и обработкой информации по отказам высокотехнологичного оборудования, а также с дальнейшем обеспечением работ по его ремонту, восстановлению, взаимодействию с поставщиками и потребителями [1].

Вместе с тем, малые и средние сервисные организации, связанные с полупромышленным оборудованием или бытовой техникой, часто не обладают большими финансовыми возможностями, для обработки поступающей информации по отказам и аварийным ситуациям. В таких компаниях часто создаются банки данных и базовая статистическая аналитика. При этом, там же обычно не хватает средств и мощностей обрабатывать исходные данные более продвинутыми способами, чем рядовыми методами статистики. Эта проблема приводит к нерациональным решениям по ремонту и техническому обслуживанию того или иного бытового и полупромышленного энергетического оборудования, перерасходу материалов и комплектующих, транслированию неполной или неточной информации производителям и потребителям о причинах отказов и сбоев [2]. В связи с этим возникает задача разработать такой программный пакет, который будет позволять проводить предварительную подготовку статистических данных и далее комплексно анализировать причины возникновения отказов, устанавливать взаимосвязь между ними, определять динамику во времени и их частотную зависимость от выполняемых работ по обслуживанию и ремонту.

Для этого можно использовать как имеющиеся платформы и программные продукты по обработке статистических данных, так и собственные разработки. В настоящем исследовании представим подход по обработке массива статистических данных по отказам методом кластерного анализа при помощи, существующей библиотеки scikit-learn на интерактивной вычислительной среде Jupyter Notebook и собственной программной разработки, а также оценим доступность обоих программных решений для рядового пользователя и эффективность анализа данных.

Отказы двухконтурных котлов. Группировка и предварительная обработка статистической информации

В ходе детальной оценки технического состояния оборудования как правило, на начальном этапе определяются со стратегией прогноза и прогностическим фоном, а далее разрабатывают систему параметров, отражающую характер и структуру технического

объекта. Далее происходит разработка нормативной и поисковых моделей оборудования, производиться моделирование и оценивается точность построения моделей (скорость выхода из строя оборудования, соответствие начальных и конечных параметров). Модели верифицируются на основе статистики, экспериментальных данных или экспертного метода. После чего происходит выработка рекомендаций для оптимизации принятия решения в планировании и управлении эксплуатацией оборудования на основе полученных моделей.

Однако такой путь, как правило, приемлем либо для разработчиков оборудования (которые устанавливают гарантию на его безаварийный срок работы), либо для крупных сервисных центров, отвечающих за высокотехнологичные промышленные аппараты. Малым и средним сервисам занимающих основную долю рынка по ремонту полупромышленного и бытового оборудования для оперативной обработки статистики по отказам выгоднее использовать наиболее информативные параметры, позволяющие оценить техническое состояние оборудования комплексно и при этом быстро.

В случае с двухконтурными газовыми котлами в исследовании были отобраны данные по наработке котелкового оборудования на отказ, коэффициенту полезного действия двухконтурных котлов и по тяжести отказа, предшествовавшего ремонту [3]. В целом отметим, что любой комплекс параметров технического состояния может быть использован для анализа, однако наибольшая информативность будет получена если обрабатываемые данные обладают полнотой (в полной мере отражают результаты отказов оборудования), достоверностью (проверяемы экспериментально), точностью, целостностью, взаимосвязанностью, измеримостью и при этом способностью отражать разные технико-технологические характеристики объекта во времени [4].

Данные, использованные для исследования (наработка котелкового оборудования на отказ, коэффициент полезного действия котла и тяжесть отказа, предшествовавшего ремонту) обрабатывались и подготавливались для исследования поэтапно. Выборка при исследовании формировалась из 4 групп двухконтурных газовых котлов марок — Ariston, Buderous, Bosh, Wisseman, по 25 котлов после отказа и последующего ремонта в каждой группе. Первым исследуемым параметром являлась наработка на отказ. Это технический параметр, характеризующий надежность восстанавливаемого прибора, устройства или технической системы, в данном случае двухконтурного газового котла [5]. Функция надежности может быть определена через вероятность безотказной работы. Формула вероятности безотказной работы выглядит так:

$$P(t) = exp^{\left(-\frac{t}{T_o}\right)}, \quad (1)$$

где t — время работы оборудования, T_o — наработка до отказа.

Наработку оборудования на отказ можно также установить через формулу интенсивности отказов [6]. Интенсивность отказов является параметром, определяющим надежность элемента:

$$\lambda = \frac{1}{\text{MTBF}}, \quad (2)$$

где λ — интенсивность отказов, а MTBF — наработка между отказами.

Но в целом наработка на отказ, определяется статистически через время:

$$T = \frac{\sum_1^m t_i}{m}, \quad (3)$$

где t_i — наработка до наступления отказа i; m — число отказов.

При этом нужно оговориться, что в настоящем исследовании речь идет о наработке до отказа, о той разновидности наработки на отказ — которая измеряется от момента первого запуска оборудования в работу — до первого выхода из строя перед ремонтом [7]. Таким образом данный параметр показывает безотказность работы нового оборудования, закупленного и используемого потребителем во времени (что крайне важно, как для потребителя — планировать свои расходы, так и для сервисной организации — планировать срок поступления заказов в зависимости от предшествовавших объемов продаж того или иного оборудования на рынке) [8].

В качестве второго исследуемого в выборке параметра был использован КПД котелкового оборудования перед выходом из строя. КПД в данном случае эффективен как аналитический параметр поскольку с одной стороны (со стороны производителя и сервисной компании) показывает динамику деградации технической системы во времени, а с другой стороны (со стороны потребителя) показывает эффективность работы оборудования до вывода в ремонт. Основная суть КПД котелкового оборудования заключается в соотношение объема потребляемого топлива к объему выделяемого тепла [9]. Рассчитать КПД можно несколькими методами. Первый метод — прямой, выражается следующим образом:

$$\eta_к = \frac{Q_к}{Q_н^p},$$

где $\eta_к$ — КПД котла; $Q_к$ — полезная энергия, переданная теплоносителю; $Q^p_н$ — тепловая энергия, выделенная в результате химической реакции горения.

Представим и обратный метод [10]. Он выглядит следующим образом:

$$Q^p_н = Q_1 + Q_2 + Q_3 + Q_4 + Q_5, \quad (5)$$

где $Q^p_н$ тепловая энергия, выделенная в результате химической реакции горения; $Q_1 = D \cdot (h_п - h_{пв}) / B$ — теплота, полезно затрачиваемая на выработку пара; D — паропроизводительность котла, кг/с; B — секундный расход топлива, кг/с или м^3/с; $h_п$ и $h_{пв}$ — энтальпия пара и питательной воды, кДж/кг; Q_2 — потери теплоты с уходящими из котельного агрегата газами; Q_3 — потери теплоты (недожог) от химической неполноты сгорания топлива; Q_4 — потери теплоты (недожог) от механической неполноты сгорания; Q_5 — потери теплоты в окружающую среду через внешние ограждения котла.

В качестве третьего анализируемого параметра была использована тяжесть последствий отказа (она была сформирована исходя из сложности ремонто-восстановительных работ котелкового оборудования и частоты возникновения). В целом тяжесть отказа может быть рассмотрена и как суммарная критичность *j*-ых отказов *i*-ых элементов оборудования (а именно сложности ремонто-восстановительных работ и той частоты, с которой они требуются) [11]. На основе Британского DEF STAN 00-60 она может быть вычислена как:

$$\sum Cm^k_{ij} = \beta^k_{ij} \cdot \alpha_{ij} \cdot \lambda_i \cdot (T_{\text{работы}})_i, \quad (6)$$

где Cm^k_{ij} — число критичности *j*-го вида отказов *i*-го элемента; β^k_{ij} — вероятность возникновения последствий определенной категории тяжести для *j*-го вида отказов *i*-го элемента; α_{ij} — доля *j*-го вида отказа *i*-го элемента; λ_i — интенсивность отказов *i*-го элемента; $(T_{\text{работы}})$ — наработка *i*-го элемента.

Визуально этот параметр может быть также представлен с помощью условной «матрицы рисков» (рис. 1).

Таким образом по всем четырем видам двухконтурных газовых котлов, по всем приведенным параметрам — были подготовлены исходные данные. Далее встала задача установить взаимосвязь между данными параметрами (технико-технологически очевидно, что подобная взаимосвязь существует) однако неизвестна ее степень, а также насколько она будет выражена по группам [12]. Для общей аналитики как уже отмечалось выгодно использовать кластерный анализ, при этом важно отметить, что

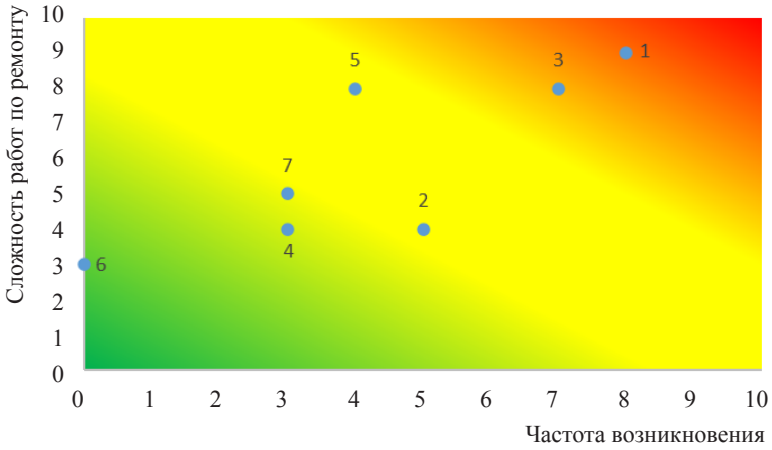

Рис. 1. Тяжесть отказа котелкового оборудования:
1 —коррозия газовых горелок; 2 — удаление продуктов нагара;
3 — выход из строя переключателя; 4 — поломка термодатчика;
5 — шум при сгорании газа; 6 — не загораются горелки;
7 — не срабатывает датчик тяги

кластеризация может быть выполнена разными способами. Тем не менее до перехода к процедуре кластеризации необходимо произвести нормировку полученных исходных данных, поскольку в их исходном виде с высокой степенью численного разброса (между КПД, наработкой на отказ и тяжестью отказа) результаты кластеризации будут хуже формализоваться и их будет сложно представить визуально.

Нормализация данных для последующего проведения кластерного анализа

Нормализация — это процесс преобразования ряда данных к одному безразмерному диапазону значений. В зависимости от цели нормализации, существуют различные функции, сопоставляющие исходным данным ряда их нормализованное значение. Среди таких функций — функция масштабирования, которая ставит в соответствие значению из диапазона $[\min(X), \max(X)]$, где $X = x_1, x_2, ..., x_n$ — ряд экспериментальных данных, значение из заданного диапазона $[a, b]$. В частном случае, когда $[a, b] = [0, 1]$, существуют несколько вариантов функции масштабирования. На пример: min-max нормализация, z-оценка и масштабирование относительно единичного вектора. Min-max нормализация, самый простой из вариантов функций масштабирования, имеет следующий вид: $f(x_i) = \frac{x_i - \min(X)}{\max(X) - \min(X)}$. Z-оценка позволяет преобразовать ряд данных к виду, имеющему нулевые значения средней арифмети-

ческой и дисперсии. Масштабирование относительно единичного вектора подразумевает представление ряда n данных в качестве n мерного вектора, после чего применяется формула нормирования вектора $X' = \frac{X}{|X|}$ по выбранной функции нормы.

В рамках текущего исследования, целью нормализации является приведение трехмерных данных, каждая компонента которых имеет разную единицу измерения, по абсолютной величине отличающиеся на порядки, к относительным безразмерным величинам. В соответствии с целью была выбрана самая простая и эффективная для данной цели функция масштабирования — min-max масштабирования [13]. Соответственно каждую компоненту трехмерных данных $A = \{(x_i, y_i, z_i), i = 1..n\}$ представляем в виде отдельного ряда данных: $X = \{x_i\}$, $Y = \{y_i\}$, $Z = \{z_i\}$, и применяем min-max нормализацию к каждому из них. В результате данного процесса, получаем нормализованные ряды данных $X' = \{f(x_i)\}$, $Y' = \{f(y_i)\}$, $Z' = \{f(z_i)\}$, после чего собираем эти ряды в единое множество, восстанавливая исходный трехмерный ряд, но с нормализованными значениями $A' = \{(f(x_i), f(y_i), f(z_i)), i = 1..n\}$. Безразмерность полученных в результате нормирования данных величин следует из формулы min-max масштабирования, из того факта, что результатом деления двух величин одинаковой размерности является величина безразмерная. В результате нормирования ряда данных, его анализ упрощается, а графическое представление зависимости его данных от значений другого ряда, особенно при различии их значений, в исходных размерностях, на порядки, приобретает более наглядный вид.

Это необходимо поскольку размерность анализируемых данных делает их плохо воспроизводимыми в графическом представлении, что неудобно для дальнейшей аналитики. В целом выполнение данной процедуры возможно для любого рода данных, используемых в кластерном анализе. Перейдем далее к кластерному анализу используя нормированные данные по котелковому оборудованию.

Проведение кластерного анализа с использованием библиотеки Scikit-learn на интерактивной вычислительной среде Jupyter notebook и программы собственной разработки

Существуют разные виды кластерного анализа: вероятностный подход, который показывает отношение каждого исследуемого объекта к одному из классов кластера; логический подход, выстраивающий дендрограмму с помощью дерева решений; иерархический подход, который предполагает наличие разного вида кластеров, объединенных в единый кластер. Для рассматриваемого случая, применим метод «к-средних» [14]. Поскольку он позво-

ляет не только качественно, но и количественно распределить события на группы. Суть метода заключается в том, что он стремится уменьшить сумму квадратичных отклонений точек кластеров от центров кластеризации [15].

Кластеризация к-методом предполагает, как правило задание изначального числа кластеров [16]. В случае если происходит построение кластеров по таким параметрам как: коэффициент полезного действия устройства, наработка на отказ до ремонта, тяжесть отказа — логично предположить, что будут формироваться три основных группы кластеров:

– ненадежные устройства (в них высока тяжесть отказа или низок коэффициент полезного действия при малой наработке на отказ, т.е. оборудование вышло из строя ранее своего гарантийного срока эксплуатации или работало до выхода из строя с заниженной производительностью);

– устройства с нормативной надежностью (в них по всем рассматриваемым параметрам выход из строя либо падение производительности произошло к гарантийному моменту и как правило не имело таких тяжелых последствий);

– устройства, показавшие высокую степень надежности, в которых тяжесть отказа была незначительна, при этом сохранялась длительность работы более чем указано в гарантии без значительной потери производительности.

Для получения более точной аналитики используем два подхода для кластеризации, общедоступный пакет «scikit-learn» в среде «Jupyter notebook», а также подготовим собственную программу и далее сравним точность кластеризации по одинаковым исходным данным, которые были предварительно нормированы. Jupyter notebook — это среда разработки, где при выполнения кода результат можно увидеть сразу [17]. Преимуществом этой среды является в том, что код можно разбить на куски и выполнить их в любом порядке. Библиотека scikit-learn — это библиотека, написанная на языке Python, включающая в себя реализацию методов кластерного и регрессионного анализа.

В случае с собственной программой (также написанной на языке Phyton), возможно более точно задать исходные параметры процесса кластеризации. Представим в виде блок схем этапы кластеризации обоими способами (рис. 2).

В первой блок схеме описан алгоритм кластеризации с помощью библиотеки scikit-learn. После начала, в первом блоке выполняется ввод данных, после чего они передаются функции k_means. Данная функция библиотеки scikit-learn производит кластеризацию по алгоритму k-means и записывает результат в переменную *X*, после чего создается пустой массив centers, предназначение которого, в последствии, хранить координаты цен-

тров кластеров. Функция silhouette_score рассчитывает средний силуэт кластеризации и записывает результат в переменную silhouette_avg. Далее в цикле рассчитываются центры каждого кластера как средняя точка в кластере с помощью функции mean, после чего записываются в массив centers. В последнем блоке, перед концом, выводятся результаты, а именно центры кластеров, индекс Данна, силуэт, отделимость каждого кластера индивидуально и общая отделимость.

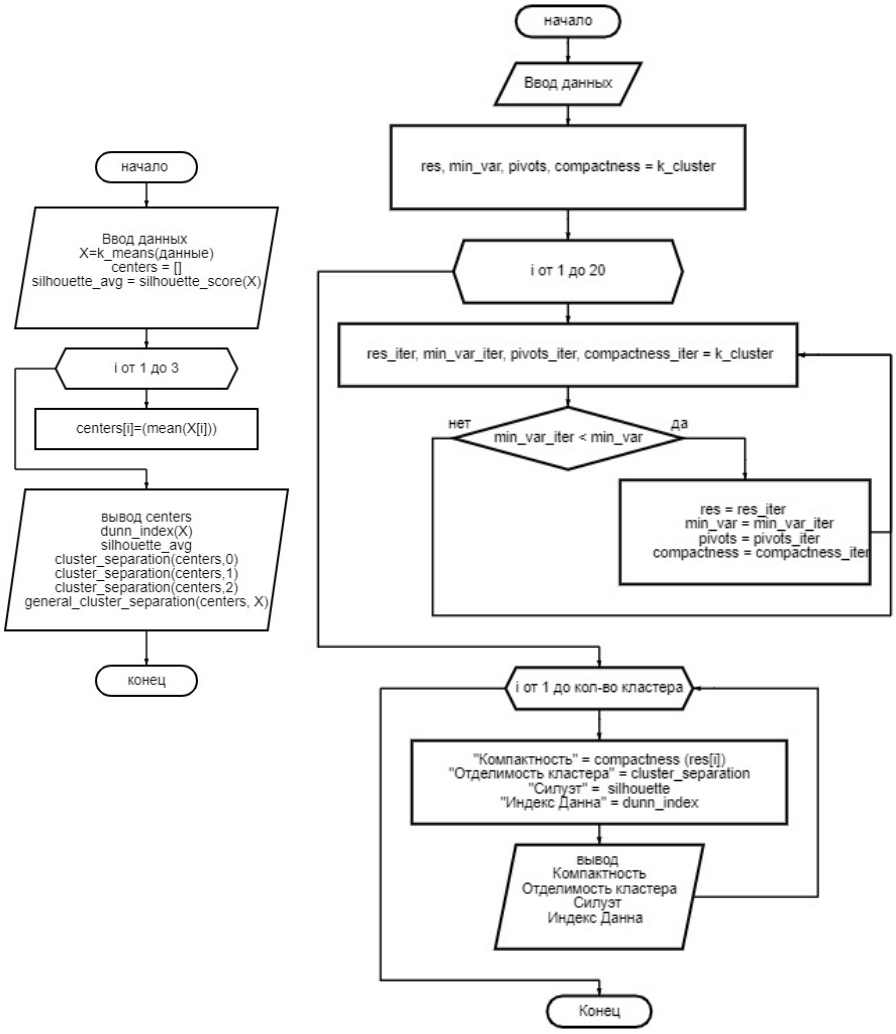

Рис. 2. Процесс кластеризации к-методом при помощи: а) библиотеки scikit-learn интерактивной вычислительной среды «Jupyter Notebook»; б) собственной программы по проведению кластерного анализа

Вторая блок схема описывает алгоритм собственной программы. Первый блок после начала — ввод данных. После ввода данных запускается алгоритм кластеризации с помощью функции k_means, которая возвращает параметры res, min_var, pivots и compactness, массив с тремя кластерами, сумма квадратичных отклонений элементов каждого кластера от его центра, центры кластером и компактность кластеров соответственно. После чего процесс повторяется в цикле двадцать раз, отбирая кластеризацию с минимальной суммой квадратичных отклонений кластеров. Выбрав лучшую кластеризацию, алгоритм далее рассчитывает для каждого кластера следующие оценки: компактность, отделимость, силуэт и индекс Данна.

На основе двух представленных алгоритмов по четырем группам нормированных данных по отказам газовых котлов марок — Ariston, Buderous, Bosh, Wisseman, проведем кластерный анализ, результаты кластеризации представим графически (рис. 3–6). Во всех случаях, как при использовании библиотеки scikit-learn в интерактивной вычислительной среде «Jupyter Notebook», так и при использовании собственной программы образовались четкие кластеры.

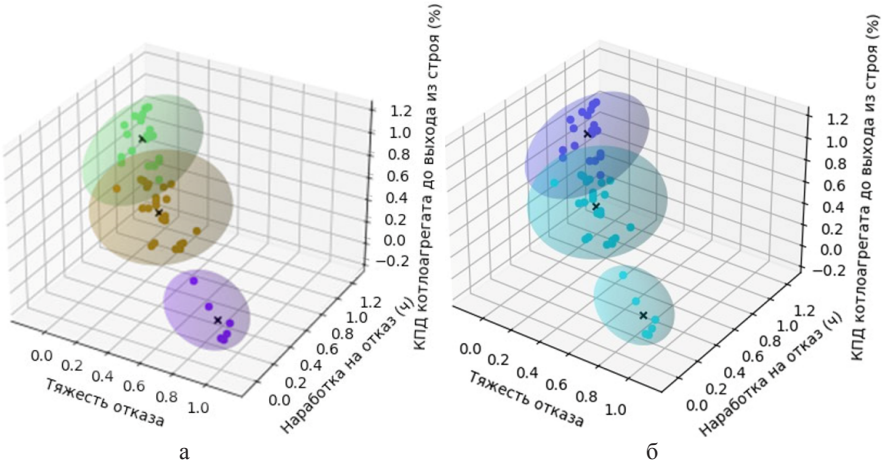

Рис. 3. Кластеризация данных по отказам котлов марки Ariston: а — с использованием библиотеки scikit-learn интерактивной вычислительной среды «Jupyter Notebook»; б — с использованием собственной программы

При этом визуально отличия в результатах кластеризации минимальны. Фиксируются центры кластеризации по каждому из кластеров, четко очерчивается группа данных. Сравнивая полученные результаты друг с другом, можно визуально отметить различия не столько между использованными алгоритмами кла-

стеризации, сколько между размерами кластеров по различным маркам газовых котлов.Так изначально предполагалось, что сформируются три основных группы кластеров: ненадежные устройства, устройства с нормативной надежностью и устройства с высокой степенью надежности. На представленных рисунках кластеры котлов с наименьшей надежностью находятся ближе всего к читателю, с нормативной надежностью в «середине куба кластеризации», и с высокой надежностью в левом верхнем углу.

Рассматривая все рисунки очевидно, что у котлов марки Ariston наибольшие по размеру кластеры сформированы как раз в группах с высокой и нормативной надежностью (рис. 3), у котлов марки Bosh (рис. 4), наибольший кластер формируется в зоне нормативной надежности, а вот у котлов марок Buderus и Viessmann (рис. 5–6) — наибольшие по размеру кластеры формируются в зоне низкой надежности.

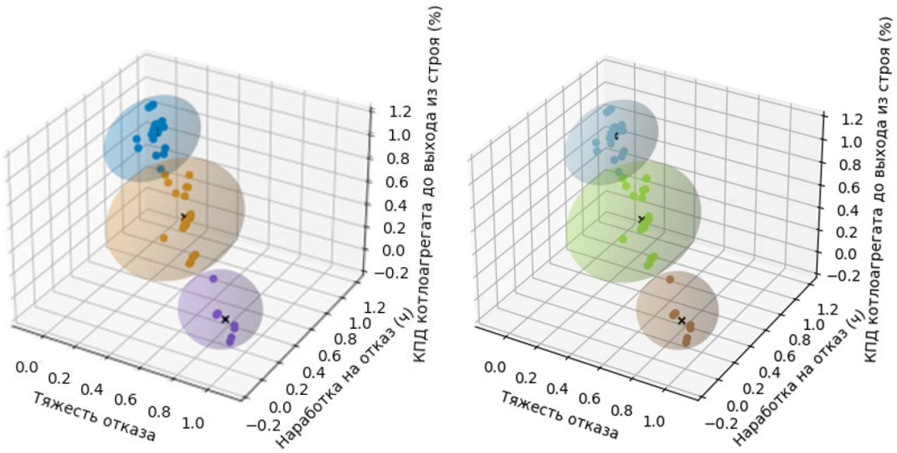

Рис. 4. Кластеризация данных по отказам котлов марки Bosh: а — с использованием библиотеки scikit-learn интерактивной вычислительной среды «Jupyter Notebook»; б — с использованием собственной программы

В целом, даже визуальная оценка результатов кластерного анализа — дает многое для технических специалистов и управленцев. В частности, возникает обусловленная необходимость изучения причин повышенной частоты и тяжести отказов котлов Buderus и Viessmann.

Возникает потребность в принятии мер по смещению статистики распределения отказов из зоны «низкой надежности» в зоны с высокой и нормативной надежностью.

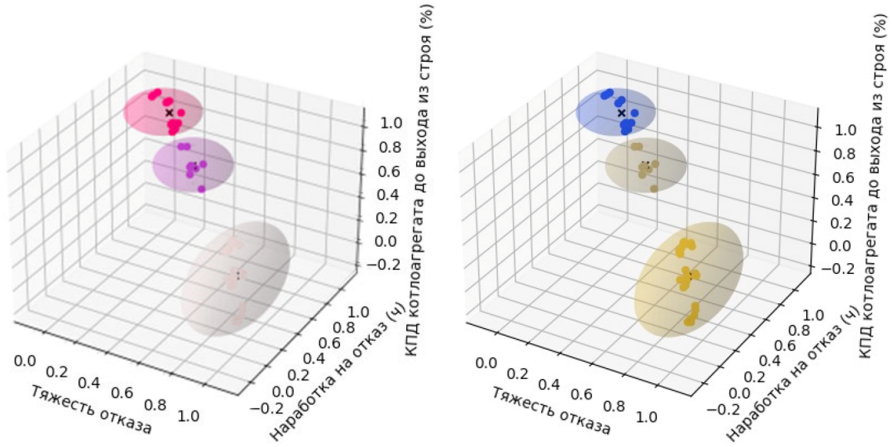

Рис. 5. Кластеризация данных по отказам котлов марки Buderus: а — с использованием библиотеки scikit-learn интерактивной вычислительной среды «Jupyter Notebook»; б — с использованием собственной программы

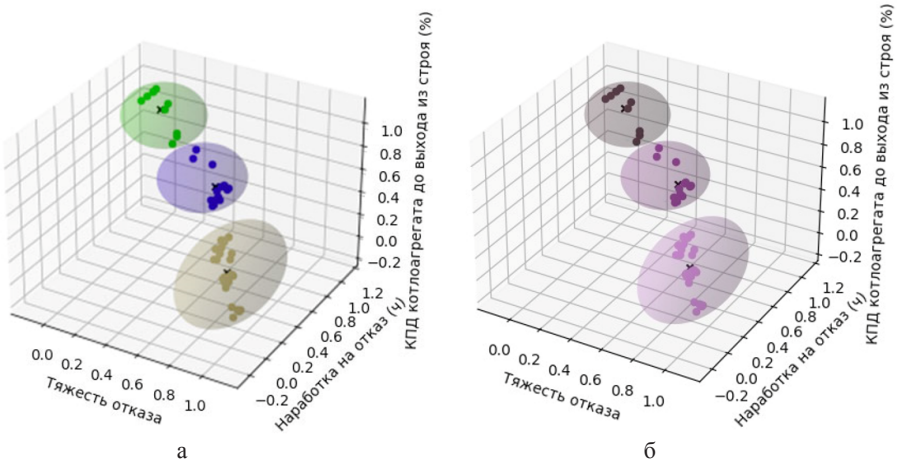

Рис. 6. Кластеризация данных по отказам котлов марки Viessmann: а — с использованием библиотеки scikit-learn интерактивной вычислительной среды «Jupyter Notebook»; б — с использованием собственной программы

Вместе с тем очевидно, что визуальной оценки недостаточно для определения эффективности кластеризации данных с помощью двух предложенных подходов. Для того, чтобы определить, что же является более точным и эффективным в обработке данных по отказам котелкового оборудования библиотека scikit-learn интерактивной вычислительной среды «Jupyter Notebook» или собственная программа рассчитаем несколько ключевых параметров

кластеризации. Они структурно укажут на плотность полученных кластеров, их отделимость друг от друга, на тесноту связей между сгруппированными событиями [18]. При этом можно будет сравнить полученные результаты по каждому из четырех массивов данных по отказам (Ariston, Buderous, Bosh, Wisseman), и по обоим примененным подходам.

Для проведения сравнения используем в частости такой параметр как индекс Дана. Индекс Данна — это метрика, по оценке алгоритмов кластеризации. При его определении цель состоит в том, чтобы идентифицировать наборы кластеров, которые являются компактными, с небольшой дисперсией между членами кластера и хорошо разделены, где средние значения разных кластеров достаточно далеки друг от друга по сравнению с дисперсией внутри кластера. При заданном распределении кластеров более высокий индекс Данна указывает на лучшую кластеризацию [19]. Его можно рассчитать различными способами, в частности:

$$D(C) = \frac{min_{Ck \in C}\left\{min_{Cl \in C \setminus Ck}\{\delta(Ck, Cl)\}\right\}}{max_{Ck \in C}\{\Delta(Ck)\}}, \tag{7}$$

где Ck, Cl — Кластеры из множества C ; C — Множество кластеров; δ —меж кластерное расстояние; Δ —диаметр кластера. В качестве функции расстояния в данной, и в всех последующих оценках, используется евклидова норма. Расстояние между кластерами считаем минимальным расстоянием между точками одного и другого кластера.

Также используем индекс силуэта, который показывает, насколько объект похож на свой кластер по сравнению с другими кластерами. Он может быть рассчитан следующим образом:

$$Sil(C) = \frac{1}{N}\sum_{Ck \in C}\sum_{x_i \in C_k}\frac{b(x_i, c_k) - a(\mathrm{x_i}, \mathrm{c_k})}{max\{a(\mathrm{x_i}, \mathrm{c_k}), b(x_i, c_k)\}}, \tag{8}$$

где x_i — элемент кластера; a, b — среднее расстояние точки x_i; N — количество элементов во множестве C.

Безусловно важным является такой параметр как отделимость кластеров. Он указывает на степень перекрытия кластеров и насколько далеко друг от друга они расположены в пространстве. И в целом подтверждает правильность изначальной гипотезы о числе рассчитываемых кластеров по их удаленности друг от друга. Отделимость рассчитывается по следующей формуле:

$$BSS = n \cdot \sum_{j=1}^{M}\left(\overline{x_j} - \overline{x}\right)^2, \tag{9}$$

где n — число кластера; M — количество кластеров; j — индекс.

Еще одним значимым параметром является компактность кластера. Она оценивает удаленность расположения элементов кластера друг от друга. Это свойство можно выразить через расстояния между элементами в кластере, плотностью внутри кластера или же объемом, занимаемым кластером в многомерном пространстве. Компактность рассчитывается как:

$$WSS = \sum_{j=1}^{M} \sum_{i=1}^{|C_j|} \left(x_{ij} - \overline{x}_j\right)^2, \quad (10)$$

Сведем полученные данные по центрам кластеров, силуэту, отделимости кластеров, компактности, индексу Дана в единую таблицу.

Сводные данные по результатам кластеризации

	Ariston		Bosch		Buderus		Viessmann	
	Jupiter	Собст. программа	Jupiter	Собст. программа	Jupiter	Собст. программа	Jupiter	Собст. программа
Индекс Данна	0.195	0.200	0.259	0.260	0.223	0.220	0.235	0.240
Индекс Силуэта	0.509	0.841	0.628	0.965	0.627	0.923	0.594	0.911
Отделимость	0.497	0.498	0.528	0.528	0.502	0.503	0.478	0.478
Компак.	2.043	2.043	1.463	1.463	1.62	1.62	1.639	1.639
Центры кластеров	0.938, 0.069, 0.113	0.938, 0.069, 0.113	0.528, 0.383, 0.609	0.528, 0.383, 0.609	0.156, 0.855, 0.951	0.156, 0.855, 0.951	0.846, 0.250, 0.172	0.846, 0.250, 0.172
	0.432, 0.354, 0.644	0.432, 0.354, 0.644	0.116, 0.817, 0.913	0.116, 0.817, 0.913	0.872, 0.256, 0.197	0.872, 0.256, 0.197	0.580, 0.619, 0.576	0.580, 0.619, 0.576
	0.106, 0.751, 0.908	0.106, 0.751, 0.908	0.937, 0.073, 0.114	0.937, 0.073, 0.114	0.431, 0.608, 0.75	0.431, 0.608, 0.750	0.100, 0.866, 0.930	0.100, 0.866, 0.930

Представленные в таблице результаты свидетельствуют, что применение собственной специальной программы позволяет несколько более точно формировать данные в кластеры. Это приобретает особое значение на большом объеме выборки. Так в большинстве случаев по всем рассматриваемым параметрам: силуэту, отделимости кластеров, компактности, индексу Дана результаты статистической обработки у собственной программы больше на сотые доли. Это говорит о более четкой организованности кластеров у собственного программного обеспечения, что очевидно связано с спецификой применяемых интерактивной вычислительной средой Jupyter Notebook кодировок [19]. В частности, с тем, что изначально библиотека scikit-learn написана сделана как наиболее

универсальный инструмент по математическому анализу данных и в применении отдельно взятых методов обработки информации может давать чуть большие отклонения [20].

Можно утверждать, что для проведения аналитики небольших выборок с заданной точностью до сотых единиц возможно использовать кластеризацию применяя библиотеки scikit-learn интерактивной вычислительной среды Jupyter Notebook, в иных случаях для исследования больших объемов статистических данных при оценке качества технического обслуживания и ремонта — выгоднее прописывать собственное программное обеспечение по кластерному анализу.

Выводы

Полученные результаты позволяют утверждать, что оба подхода как с использованием общедоступного программного обеспечения scikit-learn в среде Jupyter Notebook, так и собственной специализированной программы позволяют проводить кластеризацию данных по статистике отказов с высокой точностью. Выбор за конкретным способом (использование имеющегося программного обеспечения или написание собственного) остается непосредственно за пользователем.

В целом обеими программными подходами подтверждаются полученные результаты по отказам котелкового оборудования. Установлено, что для исследуемых выборок по 25 единиц оборудования каждой из марок (Ariston, Buderous, Bosh, Wisseman) четко формируются три группы кластеров — устройства с повышенной надежностью, устройства с нормативной надежностью и устройства с пониженной надежностью. При этом у котлов марок Buderus и Viessmann чаще встречается пониженная надежность, т.е. либо фиксируются более тяжелые отказы оборудования, либо срок его эксплуатации (до отказа) меньше установленного норматива. Наиболее надежными показали себя котлы марки Ariston — поскольку основная выборка, в обоих методах кластеризации, была смещена именно в сторону кластера повышенной надежности.

Полученные результаты позволяют внести целый ряд важных организационных, управленческих и технико-технологических решений в деятельность сервисных центров по обслуживанию двухконтурных газовых котлов. Пересмотреть сроки и объемы закупки запасных частей, переформировать графики работ по техническому обслуживанию согласно полученным результатам, провести дополнительную работу с поставщиками оборудования.

По итогам исследования получена удобная методика обработки статистической информации при помощи кластерного анализа с возможностью верификации результатов двумя программными продуктами, визуальным и численным получением резуль-

татов. Представленный концептуальный подход послужит хорошей базой для проведения более глубоких исследований надежности котелкового оборудования. В будущем резонным является добавление в обозначенную программно-аналитическую среду таких методов анализа как HAZOP, а возможно и включения предложенного подхода в комплексный цикл обеспечения надежности оборудования на основе петли качества Деминга-Шухарта. Подобные шаги позволят в случае развития программного продукта не только отслеживать послеремонтные характеристики котелкового оборудования, оценивать и анализировать отклонения в показателях надежности — но и управлять ими.

Список использованной литературы

1. Liu S. Corrosion Failure Analysis of the Heat Exchanger in a Hot Water Heating Boiler / S. Liu // Engineering Failure Analysis. — 2022. — Vol. 142. — P. 106847.

2. Evaluation of Gas-Fired Combi Boilers with HF-AHP-MULTIMOORA / F. Samanlioglu, Z. Ayağ, G. Kirkil, E. Yucal. — DOI 10.1155/2022/9225491 // Applied Computational Intelligence and Soft Computing. — 2022. — Vol. 2022. — P. 1–16.

3. Modelling of a Gas-Fired Heating Boiler Unit for Residential Buildings Based on Publicly Available Test Data / K. Simic, I. T'Jollyn, W. Faes [ed al.]. — DOI 10.1016/j.enbuild.2021.111451 // Energy and Buildings. — 2021. — Vol. 253. — P. 111451.

4. A Probabilistic Approach to Include the Overall Efficiency of Gas-Fired Heating Systems in urban Building Energy Modelling / K. Ritosa, I.D. Jaeger, D. Saelens, S. Roels. — DOI 10.1088/1742-6596/2069/1/012105 // Journal of Physics: Conference Series. IOP Publishing. — 2021. — Vol. 2069, No. 1. — P. 012105.

5. Statistics and Analysis of Test Data of Industrial Boiler Approved Products in 2020 / Y. Chang, J. Li, X. Liu [et al.] // International Conference on Advanced Manufacturing Technology and Manufacturing Systems (ICAMTMS 2022), Shijiazhuang, 2022. — Shijiazhuang. — Vol. 12309. — P. 604–610.

6. Prabhu V. Machine Learning Enabled Condition Monitoring Models for Predictive Maintenance of Boilers / V. Prabhu, D. Chaudhary. — DOI 10.1109/RDCAPE52977.2021.9633534 // RDCAPE : 2021 4th International Conference on Recent Developments in Control, Automation & Power Engineering (RDCAPE), Noida, 7 Oct. 2021. — Noida, 2021. — P. 426–430.

7. Aikin A.R. The Process of Effective Predictive Maintenance / A.R. Aikin // Tribology & Lubrication Technology. — 2021. — Vol. 77, No. 2. — P. 34–40.

8. Mohanty S. Implementation of Total Productive Maintenance (TPM) in the Manufacturing Industry for Improving Production Effectiveness / S. Mohanty, K.C. Rath, O.P. Jena // Industrial Transformation. — Boca Raton, 2022. — P. 45–60.

9. Analysis of the Resource-Saving Method for Calculating the Heat Balance of the Installation of Hot-Water Heating Boilers / N. Xodjiev, Sh. Juraev, K. Kurbanov [et al.] // AIP Conference Proceedings, 2022. — Vol. 2432, No. 1. — P. 020019.

10. Energy Efficiency Indicators for Combined Cooling, Heating and Power Systems / W. Ma, J. Fan, S. Fang, N.M.S. Hassan. — DOI 10.1016/j.enconman.2021.114187 // Energy Conversion and Management. — 2021. — Vol. 239. — P. 114187.

11. Larbi R.M. Maintenance Policies with Minimal Repair and Replacement on Failures: Analysis and Comparison / R.M. Larbi, D Aït-Kadi. — DOI 10.1080/00207543.2020.1832275 // International Journal of Production Research. — 2021. — Vol. 59, No. 23. — P. 6995–7017.

12. Leukel J. Adoption of Machine Learning Technology for Failure Prediction in Industrial Maintenance: A Systematic Review / J. Leukel, J. González, M. Riekert. — DOI 10.1016/j.jmsy.2021.08.012 // Journal of Manufacturing Systems. — 2021. — Vol. 61, No. 2. — P. 87–96.

13. Acuña-Soto C. Normalization in TOPSIS-Based Approaches with Data of Different Nature: Application to the Ranking of Mathematical Videos / C. Acuña-Soto, V. Liern, B. Pérez-Gladish. — DOI 10.1007/s10479-018-2945-5 // Annals of operations research. — 2021. — Vol. 296, No. 1. — P. 541–569.

14. Chabane A. Dependability Analysis in Systems Engineering Approach Using the FMECA Extracted From the SysML and Failure Modes Classification By K-Means / A. Chabane, S. Adjerid, I. Meddour. — DOI 10.1007/s40435-021-00855-8 // International Journal of Dynamics and Control. — 2022. — Vol. 10, No. 1. — P. 981–998.

15. Clustering Methods for Power Quality Measurements in Virtual Power Plant / F.F. Aksan, M. Jasiński, T. Sikorski [et al.]. — DOI 10.3390/en14185902 // Energies. — 2021. — Vol. 14, No. 18. — P. 5902.

16. Analysis and Research on Enterprise Resumption of Work and Production Based on K-Means Clustering / W. Peiyi, M. Longfei, L. Xianglong [ed al.]. — DOI 10.1109/ICBDA51983.2021.9403217 // 2021 IEEE 6th International Conference on Big Data Analytics (ICBDA). — Xiamen, 2021. — P. 169–174.

17. Quaranta L. KGTorrent: A dataset of Python Jupyter Notebooks from Kaggle / L. Quaranta, F. Calefato, F. Lanubile. — DOI 10.1109/MSR52588.2021.00072 // 2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR). — Madrid, 2021. — P. 550–554.

18. Fangohr H. Jupyter in Computational Science / H. Fangohr, T. Kluyver, M. DiPierro. — DOI 10.1109/MCSE.2021.3059494 // Computing in Science & Engineering. — 2021. — Vol. 23, No. 2. — P. 5–6.

19. Weiss C.J. Introducing Students to Scientific Computing in the Laboratory through Python and Jupyter Notebooks / C.J Weiss, A. Klose // Teaching Programming across the Chemistry Curriculum. American Chemical Society, 2021. — P. 57–67.

20. Ono J.P. Interactive Data Visualization in Jupyter Notebooks / J.P. Ono, J. Freire, C.T. Silva. — DOI 10.1109/MCSE.2021.3052619 // Computing in Science & Engineering. — 2021. — Vol. 23, No. 2. — P. 99–106.

References

1. Liu S. Corrosion Failure Analysis of the Heat Exchanger in a Hot Water Heating Boiler. *Engineering Failure Analysis*, 2022, vol. 124, pp. 106847.

2. Samanlioglu F., Ayağ Z., Kirkil G., Yucal E. Evaluation of Gas-Fired Combi Boilers with HF-AHP-MULTIMOORA. *Applied Computational Intelligence and Soft Computing*, 2022, vol. 2022, pp. 1–16. DOI: 10.1155/2022/9225491.

3. Simic K., T'Jollyn I., Faes W., Bastero J.B., Laverge J. Modelling of a Gas-Fired Heating Boiler Unit for Residential Buildings Based on Publicly Available Test Data. *Energy and Buildings*, 2021, vol. 253. pp. 111451. DOI: 10.1016/j.enbuild.2021.111451.

4. Ritosa K., Jaeger I.D., Saelens D., Roels S. A Probabilistic Approach to Include the Overall Efficiency of Gas-Fired Heating Systems in urban Building Energy Modelling. *Journal of Physics: Conference Series. IOP Publishing*, 2021, vol. 2069, no. 7, pp. 012105. DOI: 10.1088/1742-6596/2069/1/012105.

5. Chang Y., Li J., Liu X., Liu C., Han L. Statistics and Analysis of Test Data of Industrial Boiler Approved Products in 2020. *International Conference on Advanced Manufacturing Technology and Manufacturing Systems (ICAMTMS 2022)*, Shijiazhuang, 2022. Shijiazhuang, 2022, vol. 12309, pp. 604–610.

6. Prabhu V., Chaudhary D. Machine Learning Enabled Condition Monitoring Models for Predictive Maintenance of Boilers. *RDCAPE : 2021 4th International Conference on Recent Developments in Control, Automation & Power Engineering (RDCAPE), Noida, 7 October. 2021*. Noida, 2021, pp. 426–430. DOI: 10.1109/RDCAPE52977.2021.9633534.

7. Aikin A.R. The Process of Effective Predictive Maintenance. *Tribology & Lubrication Technology*, 2021, vol. 77, no. 2, pp. 34–40.

8. Mohanty S., Rath K.C., Jena O.P. Implementation of Total Productive Maintenance (TPM) in the Manufacturing Industry for Improving Production Effectiveness. *Industrial Transformation*. Boca Raton, 2022. pp. 45–60.

9. Xodjiev N., Juraev Sh., Kurbanov K., Sultonov S., Axatov D. Analysis of the Resource-Saving Method for Calculating the Heat Balance of the Installation of Hot-Water Heating Boilers. *AIP Conference Proceedings*, 2022, vol. 2432, no. 1, pp. 020019.

10. Ma W., Fan J., Fang S., Hassan N.M.S. Energy Efficiency Indicators for Combined Cooling, Heating and Power Systems. *Energy Conversion and Management*, 2021, vol. 239, pp. 114187. DOI: 10.1016/j.enconman.2021.114187.

11. Larbi R.M., Aït-Kadi D. Maintenance Policies with Minimal Repair and replacement on Failures: Analysis and Comparison. *International Journal of Production Research*, 2021, vol. 59, no. 23, pp. 6995–7017. DOI: 10.1080/00207543.2020.1832275.

12. Leukel J., González J., Riekert M. Adoption of Machine Learning Technology for Failure Prediction in Industrial Maintenance: A Systematic Review. *Journal of Manufacturing Systems*, 2012, vol. 61, no. 2, pp. 87–96. DOI: 10.1016/j.jmsy.2021.08.012.

13. Acuña-Soto C., Liern V., Pérez-Gladish B. Normalization in TOPSIS-Based Approaches with Data of Different Nature: Application to the Ranking of Mathematical Videos. *Annals of operations research*, 2021, vol. 296, no. 1, pp. 541–569. DOI: 10.1007/s10479-018-2945-5.

14. Chabane A., Adjerid S., Meddour I. Dependability Analysis in Systems Engineering Approach Using the FMECA Extracted From the SysML and Failure Modes Classification By K-Means. *International Journal of Dynamics and Control*, 2022, vol. 10, no. 1, pp. 981–998. DOI: 10.1007/s40435-021-00855-8.

15. Aksan F.F., Jasiński M., Sikorski T., Kaczorowska D., Rezmer J. Clustering Methods for Power Quality Measurements in Virtual Power Plant. Energies, 2021, vol. 14, no. 18, pp. 5902. DOI: 10.3390/en14185902.

16. Peiyi W., Longfei M., Xianglong L., Zhang Lu, Qin H. 16. Analysis and Research on Enterprise Resumption of Work and Production Based on K-Means Clustering. *2021 IEEE 6th International Conference on Big Data Analytics (ICBDA)*. Xiamen, 2021, pp. 169–174. DOI: 10.1109/ICBDA51983.2021.9403217.

17. Quaranta L., Calefato, Lanubile F. KGTorrent: A dataset of Python Jupyter Notebooks from Kaggle. *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)*. Madrid, 2021, pp. 550–554. DOI: 10.1109/MSR52588.2021.00072.

18. Fangohr H., Kluyver T., DiPierro M. Jupyter in Computational Science. *Computing in Science & Engineering*, 2021, vol. 23, no. 2, pp. 5–6. DOI: 10.1109/MCSE.2021.3059494.

19. Weiss C.J., Klose A. Introducing Students to Scientific Computing in the Laboratory through Python and Jupyter Notebooks. *Teaching Programming across the Chemistry Curriculum. American Chemical Society*, 2021, pp. 57–67.

20. Ono J.P., Freire J., Silva C.T. Interactive Data Visualization in Jupyter Notebooks. *Computing in Science & Engineering*, 2021, vol. 23, no. 2, pp. 99–106. DOI: 10.1109/MCSE.2021.3052619.

Информация об авторах

Щербань Павел Сергеевич — кандидат технических наук, доцент отраслевого научного кластера «Институт высоких технологий», Балтийский федеральный университет им. И. Канта, г. Калининград, e-mail: ursa-maior@yandex.ru.

Абу-Хамди Реда Валидович — студент, отраслевой научный кластер «Институт высоких технологий», Балтийский федеральный университет им. И. Канта, г. Калининград, e-mail: rabouhamdi@gmail.com.

Карагадян Артур Наириевич — студент, отраслевой научный кластер «Институт высоких технологий», Балтийский федеральный университет им. И. Канта, г. Калининград, e-mail: a.karagadian2001@gmail.com.

Information about the Authors

Pavel S. Shcherban — PhD in Technical Sciences, Associate Professor of the Branch Scientific Cluster “Institute of High Technologies”, Immanuel Kant Baltic Federal University, Kaliningrad, Russian Federation, e-mail: ursa-maior@yandex.ru.

Reda V. Abu-Hamdi — Student, Branch Scientific Cluster “Institute of High Technologies”, Immanuel Kant Baltic Federal University, Kaliningrad, Russian Federation, e-mail: rabouhamdi@gmail.com.

Artur N. Karagadyan — Student, Branch Scientific Cluster “Institute of High Technologies”, Immanuel Kant Baltic Federal University, Kaliningrad, Russian Federation, e-mail: a.karagadian2001@gmail.com.

Вклад авторов

Все авторы сделали эквивалентный вклад в подготовку публикации. Авторы заявляют об отсутствии конфликта интересов.

Contribution of the Authors

The authors contributed equally to this article. The authors declare no conflicts of interests.

Для цитирования

Щербань П.С. Использование библиотеки Scikit-learn и собственной программы по кластеризации для обработки статистических данных по отказам двухконтурных газовых котлов / П.С. Щербань, Р.В. Абу-Хамди, А.Н. Карагадян. — DOI 10.17150/2713-1734.2023.5(2).172-191. — EDN VDHPRA // System Analysis & Mathematical Modeling. — 2023. — Т. 5, № 2. — С. 172–191.

For Citation

Shcherban P.S., Abu-Hamdi R.V., Karagadian A.N. Using the Scikit-Learn Library and Our Own Clustering Program to Process Statistical Data on Failures of Double-Circuit Gas Boilers. *System Analysis & Mathematical Modeling*, 2023, vol. 5, no. 2, pp. 172–191. (In Russian). EDN: VDHPRA. DOI: 10.17150/2713-1734.2023.5(2).172-191.