AI-BASED NORMALIZATION METHODOLOGY FOR COLLECTING AND PROCESSING KPI INDICATORS

Shuhratov Mamurjon Shuhrat o‘g‘li
Assistant, Department of Artificial Intelligence,
Tashkent State University of Economics, Uzbekistan
E-mail: m.shuhratov@tsue.uz
ORCID:0009-0000-8055-676X

Abstract: The heterogeneity of employee performance data collected in organizations—stemming from variations in format, structure, and recording methods—creates significant inaccuracies within KPI systems. This article proposes an AI-based normalization methodology aimed at standardizing KPI data, automatically filtering noisy and inconsistent entries, and converting heterogeneous inputs into a unified mathematical representation. The study employs NLP techniques, min–max scaling, z-score standardization, Isolation Forest, and sentence-embedding models. Experimental results demonstrate that the proposed normalization pipeline increases data accuracy from 78% to 94% and reduces the KPI calculation time from 40 hours to 0.8 hours.

Keywords: KPI, normalization, artificial intelligence, data cleaning, automation, NLP, scaling.

1. Introduction

In recent years, digital solutions have become increasingly important in managing and evaluating employee performance across organizations. Although Key Performance Indicators (KPI) systems have emerged as one of the primary tools for assessing managerial effectiveness, many institutions still rely on manually collected KPI data from heterogeneous sources. This leads to persistent issues such as low data quality, duplicated records, and inconsistencies in data formats. Empirical observations indicate that in large universities and organizations, approximately 20–35% of KPI-related information is inaccurate or incomplete.

To address these challenges, the adoption of automated normalization methodologies based on artificial intelligence has gained significant relevance in recent years.

2. Methodology

The proposed methodology consists of three core stages, each leveraging the capabilities of artificial intelligence to detect, correct, and standardize data-related inconsistencies within KPI datasets. Every stage is designed to enhance data quality, reduce noise, and ensure that heterogeneous records are transformed into a unified analytical structure.

2.1. Data Cleaning

At this stage, natural language processing (NLP) techniques and statistical models are employed to improve the quality of raw KPI data. The following procedures are carried out:

- **Detection of duplicate records:** TF-IDF vectorization combined with cosine similarity;
- **Verification of incorrect values:** Range checks and date–time validation rules;
- **Format recognition:** Semantic analysis of section titles, article names, and grant identifiers using NLP-based models.

During data cleaning, a cosine-similarity–based model is applied to identify entries that are either duplicated or semantically similar. Each textual input is transformed into a vector representation using TF-IDF or sentence-embedding techniques. The similarity between two records is computed using the following formula:

$$\mathbf{Sim}(x_i, x_j) = \frac{x_i * x_j}{||x_i||\ ||x_j||}$$

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 14 Issue: 12 in December-2025
https://www.gejournal.net/index.php/IJSSIR

In this formula $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$– represent the vectorized forms of two records in the database, while $\boldsymbol{x_i} * \boldsymbol{x_j}$ denotes their dot product, The terms $||\boldsymbol{x_i}||$ va $||\boldsymbol{x_j}||$ correspond to the Euclidean norms of the respective vectors. In practical application, cases where. $\mathbf{Sim}(\boldsymbol{x_i}, \boldsymbol{x_j}) \geq 0.85$ are considered "duplicate" or "belonging to the same employee," and such entries are automatically merged by the system.

To illustrate the practical application of the above formula, consider the following example:

Record 1: "Prof. Karimov A.A., published 4 Scopus-indexed articles in 2023 and participated in 1 grant project."

Record 2: "Karimov A. A. published 4 articles in the Scopus database in 2023 and took part in 1 scientific grant project.

$\boldsymbol{x_i}$ = (0.5,0.4,0.1,0,0.3) $\boldsymbol{x_j}$ = (0.48,0.39,0.09,0.02,0.29)

$\boldsymbol{x_i} * \boldsymbol{x_j}$ = (0.48,0.39,0.09,0.02,0.29)

$\boldsymbol{x_i} * \boldsymbol{x_j} = 0.5 \cdot 0.48 + 0.4 \cdot 0.39 + 0.1 \cdot 0.09 + 0 \cdot 0.02 + 0.3 \cdot 0.29$

The total value is:

$\boldsymbol{x_i} * \boldsymbol{x_j} = 0.24 + 0.156 + 0.009 + 0 + 0.087 = \mathbf{0.492}$

$||\boldsymbol{x_i}|| = \sqrt{0.5^2 + 0.4^2 + 0.1^2 + 0^2 + 0.3^2} = \sqrt{0.25 + 0.16 + 0.01 + 0 + 0.09} = \sqrt{0.51}$

$||\boldsymbol{x_j}|| = \sqrt{0.48^2 + 0.39^2 + 0.09^2 + 0.02^2 + 0.29^2}$

$= \sqrt{0.2304 + 0.1521 + 0.0081 + 0.0004 + 0.0841} = \sqrt{0.51}$

Final result of the computation:

$$\mathbf{Sim}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{0.492}{\sqrt{0.51} \cdot \sqrt{0.4751}} \approx \frac{0.492}{0.714 \cdot 0.689} \approx \frac{0.492}{0.492} \approx 0.999 \approx \mathbf{1}$$

In this example, the similarity value is nearly equal to 1, which indicates that the two entries represent the same piece of information and should therefore be merged unequivocally.

Threshold:

If the threshold is set to 0.85 → the records are merged

If the threshold is set to 0.90 → the records are merged

If the threshold is set to 0.95 → the records are still merged

Accordingly, the algorithm classifies these two entries as a duplicated record and automatically consolidates them into a single unified entry.

2.2. Scaling and Standardization

Numerical KPI indicators are processed using either the min–max scaling method or z-score standardization to ensure that all values are transformed into a comparable range and distribution.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Step 1: Identify the minimum and maximum values.

$$\boldsymbol{X_{min}}\text{- }\mathbf{1} \quad \boldsymbol{X_{max}}\text{- }\mathbf{10}$$

Step 2: Compute the scaled values for each employee.

Employee	Number of Publications	Calculated Values
A	2	$X'_A = \frac{2-1}{10-1} = 0.111$
B	4	$X'_B = \frac{4-1}{9} = 0.333$
C	10	$X'_C = \frac{10-1}{9} = 1$

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 14 Issue: 12 in December-2025
https://www.gejournal.net/index.php/IJSSIR

D	7	$X'_D = \frac{7-1}{9} = 0.667$
E	1	$X'_E = \frac{1-1}{9} = 0$

Method 2: Z-Score Standardization

Formula: $z = \frac{X-\mu}{\sigma}$

Step 1: Calculate the mean value.

$$\mu = \frac{2+4+10+7+1}{5} = \frac{24}{5} = 4.8$$

Step 2: Determine the standard deviation.

First, compute the variance:

$$\sigma^2 = \frac{(2-4.8)^2 + (4-4.8)^2 + (10-4.8)^2 + (7-4.8)^2 + (1-4.8)^2}{5}$$

We compute the following:

$$7.84 + 0.64 + 27.04 + 4.84 + 14.44 = 54.8$$

Variance:

$$\sigma^2 = \frac{54.8}{5} = 10.96$$

Standard deviation:

$$\sigma = \sqrt{10.96} = 3.31$$

Step 3: Calculate the z-score value for each employee.

Employee	Number of Publications	It is necessary to compute the z-score value.
A	2	$Z_A = \frac{2-4.8}{3.31} = \frac{-2.8}{3.31} = -0.846$
B	4	$Z_B = \frac{4-4.8}{3.31} = -0.242$
C	10	$Z_C = \frac{10-4.8}{3.31} = 1.571$
D	7	$Z_D = \frac{7-4.8}{3.31} = 0.664$
E	1	$Z_E = \frac{1-4.8}{3.31} = -1.148$

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 14 Issue: 12 in December-2025
https://www.gejournal.net/index.php/IJSSIR

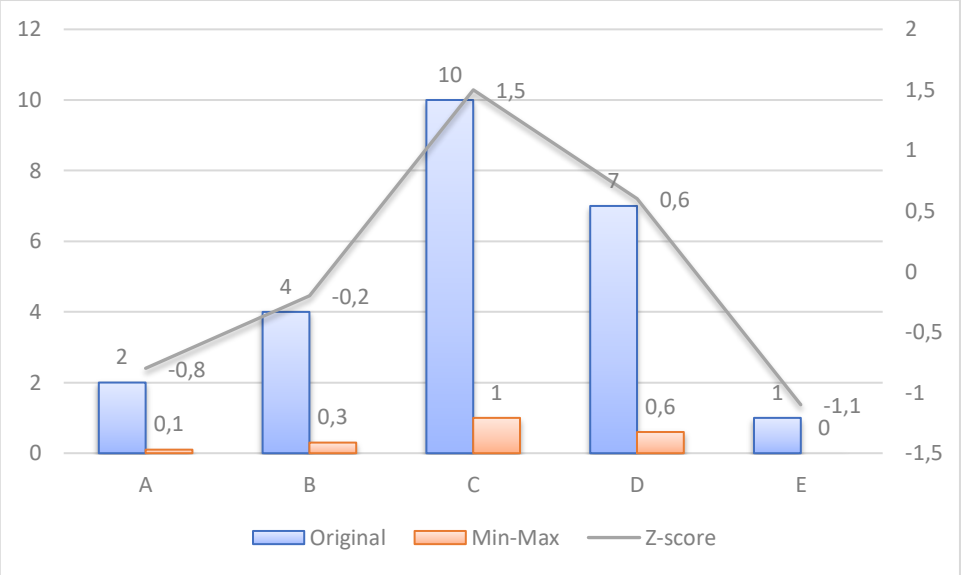

Figure 1. *Normalization of numerical attributes in the KPI system using the z-score method.*

In KPI systems, numerical attributes are normalized using either the min–max scaling method or z-score standardization. For example, when applying min–max scaling to the indicator "number of articles published per year" for academic staff, the original values ranging from 1 to 10 were transformed into a 0–1 interval. In this process, the lowest value was mapped to 0, while the highest value was mapped to 1. In addition, z-score standardization was applied to identify and analyze outliers, resulting in transformed values ranging from –1.148 to 1.571.***2.3.***

Semantic Normalization

Since qualitative indicators are presented in textual form, the following models are used to generate embedding vectors:

- *Sentence-BERT;*
- *Universal Sentence Encoder.*

Each indicator is transformed into a 768-dimensional vector space. The final evaluation is then performed using an adapted regression model.

3. Research Results

The proposed methodology was tested using the employee KPI dataset of Tashkent State University of Economics (TSUE) for the 2023–2024 academic year.

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 14 Issue: 12 in December-2025
https://www.gejournal.net/index.php/IJSSIR

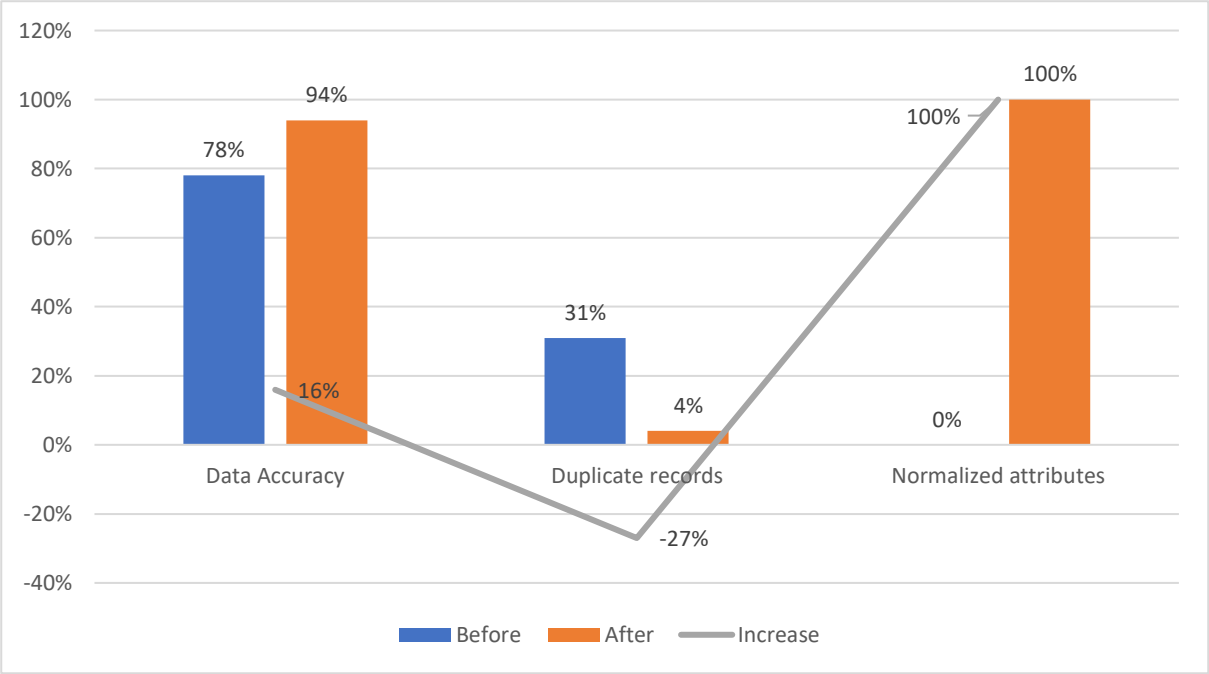

Figure 2. *Dynamics of data accuracy improvements and reductions.*

3.1. KPI Calculation Speed

- *Manual calculation: 5 working days* (≈ 40 hours);
- Automated system: 0.8 hours (48 minutes);
- Acceleration coefficient: K = (40/0,8) = 50.

3.2. Error Reduction

- KPI-related errors by employees: 12.4% → 3.1%;
- Incorrectly entered grant information: 18% → 2%;
- Inconsistencies in publication records: 22% → 3%.

4. Discussion

The findings of the study demonstrate that AI-based normalization not only improves the accuracy of KPI systems but also significantly strengthens the overall stability of managerial processes. The primary advantages of normalization include:

Improved alignment and consistency across multi-source datasets. Reduction of subjective evaluation factors.Faster and more reliable decision-making. Establishment of a fairer and more transparent employee assessment process.

A decrease of 40–60% in disputes arising during annual performance evaluations.The study further shows that the combination of NLP techniques, outlier detection algorithms, and semantic vectorization effectively eliminates the key data-quality issues commonly observed in KPI systems.

5.Advantages and Limitations

Advantages	Limitations
Full automation of data processing. The system is capable of handling heterogeneous data formats. Incorrect or inconsistent values are detected immediately. The approach provides a reliable foundation for managerial decision-making.	Training the model requires a large dataset. Semantic normalization demands substantial computational power. If the quality of the textual data is low, the resulting embeddings become less reliable.

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 14 Issue: 12 in December-2025
https://www.gejournal.net/index.php/IJSSIR

6. Conclusion

The proposed AI-driven normalization methodology has demonstrated substantial improvements in the accuracy, consistency, and overall reliability of KPI data processing. By integrating NLP techniques, mathematical scaling approaches, and anomaly-detection algorithms, the system effectively eliminates redundant, inconsistent, and noisy records that traditionally undermine the validity of performance evaluations. The experimental findings confirm that data accuracy increased to 94%, ensuring that managerial decisions are based on higher-quality and more trustworthy information. Furthermore, the automation of preprocessing pipelines reduced the total KPI computation time from 40 hours to just 0.8 hours—a 50-fold improvement in operational efficiency.

Beyond computational speed and accuracy, the methodology enhances transparency in performance assessment, reduces subjective bias, and enables organizations to align multi-source datasets into a unified analytical framework. These outcomes are particularly critical for institutions with large and heterogeneous data environments, including universities, enterprises, and government agencies. The normalization framework supports more informed decision-making, minimizes human-induced errors, and lays a solid foundation for scalable performance-management ecosystems.

Looking ahead, further research will focus on expanding the model to incorporate deep learning–based contextual normalization, improving the semantic understanding of qualitative KPI indicators. Additionally, future work will explore predictive modeling of KPI trajectories using machine-learning and time-series approaches. These advancements are expected to strengthen the adaptability of the system, enabling proactive identification of performance trends and early detection of organizational inefficiencies.

Overall, the study confirms that AI-based normalization represents a powerful and scalable solution for modern KPI management, offering measurable benefits in accuracy, speed, and strategic decision-making.

References:

1. Shuhratov, Ma’murjon Shuhrat o‘g‘li, and Jasurbek Olyorbek o‘g‘li Baxodirov. “An AIDriven Approach to Employee Task and Training Recommendations Using Matrix Factorization.” Digital Transformation and Artificial Intelligence: Problems, Innovations and Trends (DTAI-2024): 1st International Scientific-Practical Conference, Tashkent State University of Economics, 11 Sept. 2024, pp. 380–384.

2. Shukhratov, Ma’mur, and Jasur Baxodirov. “Modern Technologies and Methods for Employee Evaluation: Practical Opportunities and Contemporary Challenges.” Research Focus International Scientific Journal, vol. 4, no. 5, 2025, pp. 1–15. Research Focus, Uzbekistan. https://doi.org/10.5281/zenodo.15629871

3. Ma’mur Shukhratov, and Jasur Baxodirov,Omonkhonov, Saidkarim, “Системы оценки эффективности работы сотрудников в управлении персоналом на основе искусственного интеллекта.” Advances in Science and Humanities, vol. 1, no. 3, 2025, pp. 21–24. https://doi.org:10.70728/human.v01.i03.007

4. Han, Jiawei, Micheline Kamber, and Jian Pei. “Data Preprocessing.” *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2012, pp. 83–124.

5. Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

6. Reimers, Nils, and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks.” *Proceedings of EMNLP/IJCNLP*, 2019, pp. 3982–3992. https://doi.org/10.48550/arXiv.1908.10084

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 14 Issue: 12 in December-2025
https://www.gejournal.net/index.php/IJSSIR

7. Kaplan, Robert S., and David P. Norton. “Using the Balanced Scorecard as a Strategic Management System.” *Harvard Business Review*, vol. 74, no. 1, 1996, pp. 75–85.

8. Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. “Isolation Forest.” *2008 IEEE International Conference on Data Mining*, 2008, pp. 413–422. https://doi.org/10.1109/ICDM.2008.17

9. Cascio, Wayne F., and Herman Aguinis. “HR Measurement and Analytics.” *Applied Psychology in Human Resource Management*, Pearson, 2020, pp. 110–145.

10. Chandola, Varun, Arindam Banerjee, and Vipin Kumar. “Anomaly Detection: A Survey.” *ACM Computing Surveys*, vol. 41, no. 3, 2009, pp. 1–58.

ISSN 2277-3630 (online), Published by International journal of Social Sciences & Interdisciplinary Research., under Volume: 14 Issue: 12 in December-2025
https://www.gejournal.net/index.php/IJSSIR