UDC 004.942:614.4(460)
DOI: 10.20535/SRIT.2308-8893.2023.4.08

INFORMATION SYSTEM FOR ASSESSING THE INFORMATIVENESS OF AN EPIDEMIC PROCESS FEATURES

K. BAZILEVYCH, O. KYRYLENKO, Y. PARFENIUK, S. YAKOVLEV, S. KRIVTSOV, I. MENIAILOV, V. KUZNIETCOVA, D. CHUMACHENKO

Abstract. The primary objective of this study is to assess the informativeness of various parameters influencing epidemic processes utilizing the Shannon and Kullback–Leibler methods. These methods were selected based on their foundation in the principles of information theory and their extensive application in machine learning, statistics, and other relevant domains. A comparative analysis was performed between the results acquired from both methods, and an information system was designed to facilitate the uploading of data samples and the calculation of factor informativeness impacting the epidemic processes. The findings revealed that certain features, such as "Chronic lung disease," "Chronic kidney disease," and "Weakened immunity," did not carry significant information for further analysis and hindered the forecasting process, as per the data set examined. The developed information system efficiently supports the assessment of feature informativeness, thereby aiding in the comprehensive analysis of epidemic processes and enabling the visualization of the results. This study contributes to the current body of knowledge by providing specific examples of applying the described algorithmic models, comparing various methods and their outcomes, and developing a supportive tool for analyzing epidemic processes.

Keywords: information system, epidemic process, informativeness of features, Shannon method, Kullback–Leibler method.

INTRODUCTION

Predicting morbidity is an essential task in health care and public health. The use of machine learning in the analysis of epidemic processes is relevant in contemporary conditions, as it allows for rapid and efficient processing of large volumes of data and making accurate forecasts [1]. This helps reduce the consequences of epidemics and ensures a more effective fight against diseases. Using machine learning models helps predict morbidity with high accuracy [2].

In the modern world, especially considering the current situation related to the COVID-19 pandemic, the theme of analyzing data on epidemic processes remains extremely relevant and critically important. Data analysis is an essential tool that plays a key role and helps understand the spread of disease [3], identify trends [4], identify risk groups of the population [5], evaluate the effectiveness of control measures [6], imagine the scale of the problem [7], and predict the future development of epidemics [8]. It helps scientists, doctors, and relevant authorities make informed decisions and develop strategies for effective epidemic control [9].

It is also difficult to overestimate the importance of timely medical diagnostics in managing epidemic processes. Rapid and accurate disease diagnosis is a key factor for successful control and management of epidemics [10]. Ensuring timely diagnostics allows diagnosing and isolating sick people, starting treatment,

taking necessary preventive measures and vaccination, and taking strategic steps to reduce the spread of the disease.

Laboratory tests are one of the main tools for medical diagnostics of epidemic diseases [11]. They allow for detecting the presence of a pathogenic agent, determining its characteristic properties, and establishing a diagnosis. For example, in the case of the COVID-19 pandemic, testing for the SARS-CoV-2 virus is crucial for detecting infected individuals, even when they do not show symptoms. This helps to take appropriate control measures and preventive strategies.

Many modern healthcare facilities have information systems for storing various medical data about patients' health, used by doctors for diagnosing pathological processes [12]. However, when analyzing medical data, identifying patterns, and extracting it, one faces the problem of dimensionality. The dimensionality of stored data, determined by the number of different features describing the patient's health status, is vast and sometimes reaches several tens and hundreds of indicators [13].

Evaluating informativeness is essential for analyzing epidemic process data, as it allows for determining the significance of various factors and relationships associated with diseases [14]. This helps to identify key factors affecting the spread of epidemics and make effective decisions regarding their prevention and treatment. Informativeness evaluation also helps detect complex relationships between different factors and determine which has the most significant impact on epidemic processes [15]. This allows for making more accurate predictions and effective decisions regarding epidemic response.

Therefore, the problem of reducing the dimensionality of the feature space and identifying the most informative features is a very relevant task of epidemic process data analysis.

The aim of the paper is to develop the information system for evaluation of the factors' informativeness for healthcare data.

Research is part of a complex intelligent information system for epidemiological diagnostics, the concept of which is discussed in [16, 17].

2. MATERIALS AND METHODS

2.1. Informativeness of features

The informativeness of a feature is an indicator of its significance or usefulness for solving a specific task or problem. This is an essential concept in many areas, including machine learning, statistics, signal processing, and many others [18]. The informativeness of features is assessed depending on their ability to classify or predict the target variable. More informative features have a greater impact on the model and provide more significant information for the separation or prediction of classes.

Diagnostic features are specific symptoms, indicators, or characteristics used to diagnose a disease, condition, or problem [19]. In medicine, diagnostic features help doctors determine a disease or condition based on examination, patient surveys, laboratory tests, examinations, images, and other studies. Diagnostic features may include such indicators:

- Physical symptoms: for example, pain, pulsation, swelling, bleeding, skin color change, etc.

- Behavioral symptoms: for example, nervousness, depression, irritation, inability to concentrate, sleep change, appetite change, etc.
- Laboratory results: such as cell count, hormone level, substance concentration in the blood or urine, or results of other analyses.
- Imaging: results of X-rays, CT scans, MRI, or other techniques that may show changes in the structure or function of organs.
- Anamnesis: information obtained from the patient about their medical history, symptoms, duration, and nature of the disease.
- Genetic research: determining the presence or absence of certain genetic mutations or variants.

2.2. Problem formulation of feature space reduction

The application of modern information technologies in medicine contributes to accumulating large volumes of medical data, which are stored and processed using medical information systems (MIS). These data contain medical knowledge that can be extracted and used for decision-making, such as diagnosing pathological processes [20]. The dimensionality of the stored data, defined by the number of different features describing the patient's health status, is vast and sometimes reaches several tens and hundreds of indicators. Therefore, the problem of reducing the dimensionality of the feature space and highlighting the most informative features is very relevant for MIS development.

Let Ω be a set of objects, and $X = \{x_1, x_2, \ldots, x_n\}$ be the finite set of quantitative features of these objects. For any object $\varpi \in \Omega$, its feature description $\{x_1(\varpi), x_2(\varpi), \ldots, x_n(\varpi)\}$ is known as a n-dimensional vector, where this vector's $(i-a)$-th coordinate equals the $(i-a)$-th feature's value. The set of feature descriptions of objects for a given sample of objects $A \subseteq \Omega$ is given as a matrix of size $|A| \times n$, a table "object – feature". Let $I(Z)$ be the measure of informativeness of the subset of features $Z \subseteq X$, defined on A. It is necessary to select some subset $Z^* \subseteq X$ from all different subsets of the set X, such that

$$I(Z^*) = \max_{Z \in X} I(Z).$$

The task of features selection is computationally complex; as for $|X| = n$, a permutation of all different subsets $Z \subseteq X$ requires $O(2^n)$ time.

2.3. Kullback–Leibler Method

The Kullback–Leibler method is a statistical approach for measuring the divergence between two probability distributions. This method is popular in many fields, including statistics, machine learning, and information theory [21]. Using the Kullback–Leibler method, a measure is calculated that gauges the divergence between two distributions to assess the informativeness of a feature.

Typically, two distributions are input into the Kullback–Leibler method to evaluate the informativeness of features [22]: the distribution of data with the feature value considered and the distribution of data without considering the feature value. The method estimates the informativeness of the studied feature as a value ranging from 0 to 2. In this case, it is considered that the closer the informativeness measure $I(x)$ is to 2, the higher the informativeness of x, and conversely, the closer $I(x)$ is to 0, the lower the informativeness of x. The output of the

Kullback–Leibler method is a numerical estimate indicating the informativeness of the feature.

Algorithmic Model of the Kullback–Leibler Method

Step 1. Define the target input set (in this case, it is "Morbidity").

Step 2. Calculate the probability of the event for each value in the target set: $Q(X) = n(X)/N$, where n is the number of cases X, and N is the total number of cases.

Step 3. Calculate the probability of the event for each value in the feature: $P(y) = n(y)/N$, where n is the number of cases y, and N is the total number of cases.

Step 4. Calculate the Kullback–Leibler divergence between the two sets P and Q. The Kullback–Leibler divergence, sometimes called relative entropy, is a measure of the difference between two probability distributions:

$$D(P,Q) = \sum_i P(i)\log_2(P(i)/Q(i)),$$

where $P(i)$ is the joint probability of the event X-target set and y-feature, and $Q(i)$ is the probability of the event of the target set.

Repeat *steps 3-4* for all values in the feature and calculate the overall Kullback–Leibler divergence.

Step 5. Calculate the overall informativeness of the feature.

Step 6. Evaluate the obtained results based on the magnitude of the informativeness of the feature. The higher the evaluation value, the more informative the feature.

Step 7. Select the features with the highest values as the most informative.

The algorithm of the model is shown in Fig. 1.

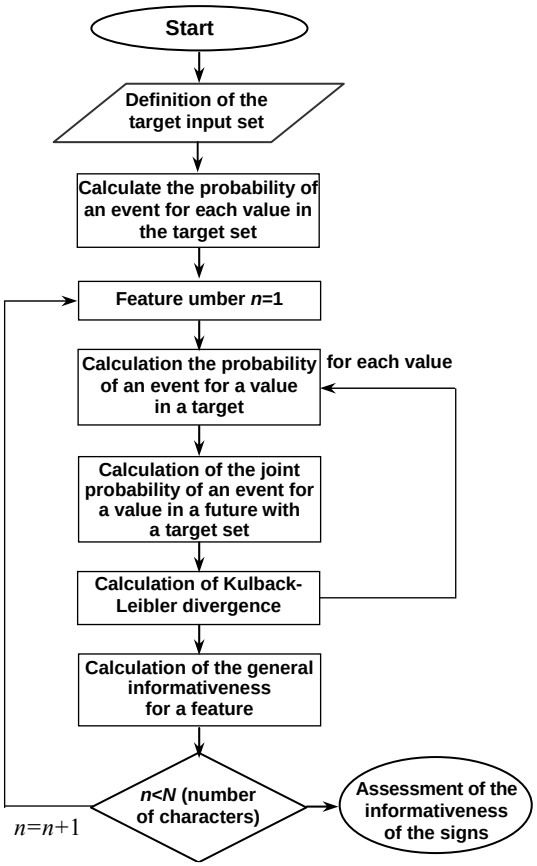


Fig. 1. The algorithm of the Kullback–Leibler method

2.4. Shannon Method

The Shannon method for calculating feature informativeness in a table is based on the concept of entropy in information theory [23]. Entropy is a measure of uncertainty or randomness in a data set. Entropy reflects the average level of 'information,' 'surprise,' or 'uncertainty' inherent in the possible outcomes of a random variable [24].

The Shannon method provides an estimate of the informativeness of the studied feature in the form of a normalized variable, which takes values from 0 to *1* [25]. In this case, the informativeness of feature x is said to be higher as $I(x)$ approaches *1* and

lower as $I(x)$ approaches 0.

Algorithmic model of the Shannon method

Step 1. Define the target input set (in our case, it is "Morbidity").

Step 2. Calculate the total entropy for the target set using the Shannon formula

$$H(S) = -\sum_{i=0}^{N} p_i \log_2 p_i ,$$

where p_i is the probability of the occurrence of the *i*-th class in the data set, H is the entropy, and S is the set of instances.

Step 3. Divide the data by each unique feature value and calculate the frequency of each value in the target set.

Step 4. Calculate the entropy for each feature value.

Step 5. Calculate the weighted entropy for each feature value, multiplying the entropy value by its frequency. Weighted entropy by the Shannon method [26] is used to measure the informational weight of a random event:

$$H_{weighed} = P(S)H(S) ,$$

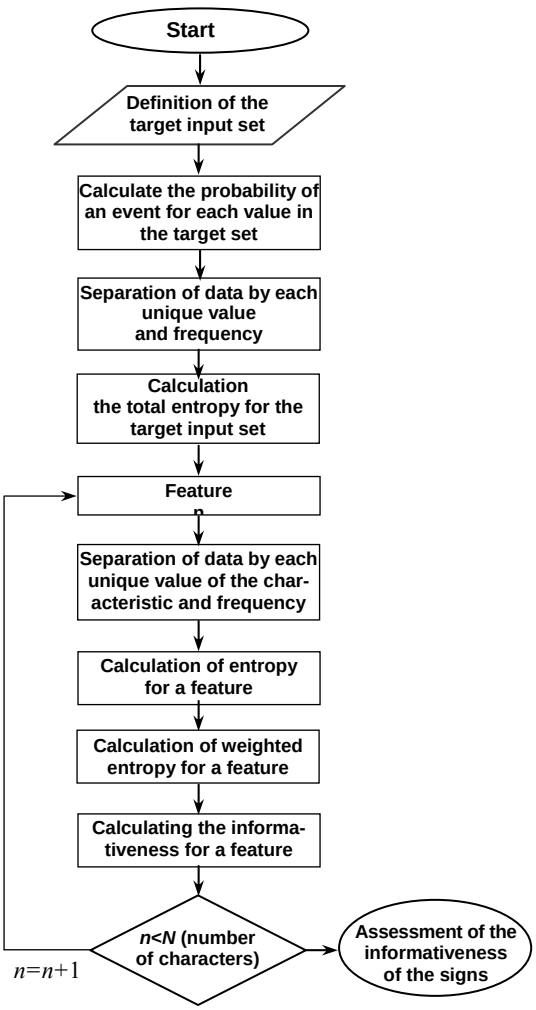


Fig. 2. The algorithm of the Shannon method

where $P(S) = m / N$: m is the frequency of the occurrence of the value in the feature; N is the total number; $P(S)$ is the probability of the occurrence of the S-th class relative to the target variable.

Step 6. Calculate the informativeness of features. The informativeness of a feature is calculated as the difference between the entropy of the output set and the sum of the entropy of the subsets formed by the given feature, with weights equal to the fraction of the subset in the output set:

$$I(S) = H(S) - \sum_{i=0}^{N} H_{weighed} ,$$

where $I(S)$ is the informativeness of the feature of the subset S.

Repeat *steps 2-6* for all features and calculate the informativeness for each feature.

Step 7. Evaluate the obtained results based on the informativeness of the feature. The higher the evaluation value, the more informative the feature.

Step 8. Select features with the highest values as the most informative.

Figure 2 shows the flowchart of the algorithmic model.

3. RESULTS

3.1. Program realization

Various algorithms and methods were employed to develop the information system, and *Python* is an ideal choice for such tasks. Its library, *sklearn*, includes many machine learning algorithms, including naive Bayes, logistic regression, and gradient boosting [27].

For *data visualization, tkinter, matplotlib.pyplot,* and *seaborn* were used, which are powerful visualization tools in Python. These libraries provide many possibilities for creating plots, diagrams, interactive visualizations, and more.

Based on data from *healthcare facilities*, the developed software product predicts the probability of a patient getting sick. The product is a decision-support system for general practitioners, which is especially important during pandemics and other disasters that limit the number of doctors.

Figure 3 shows the interface of the software product.

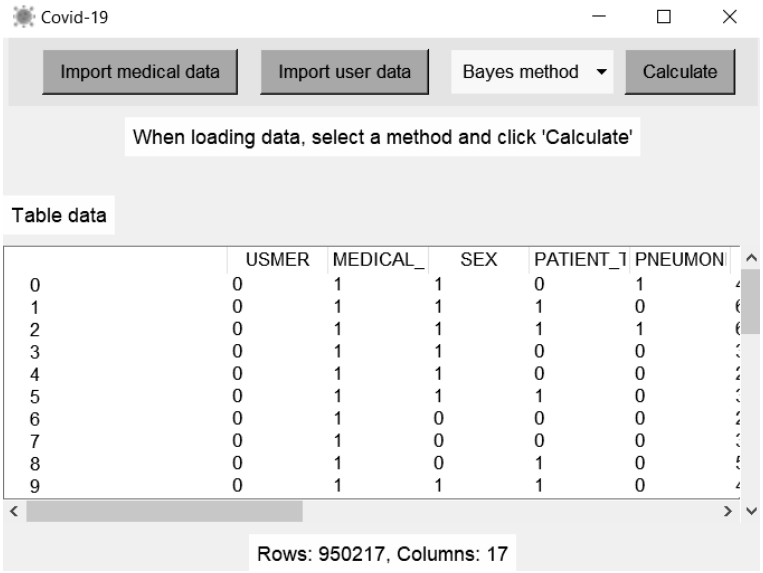


Fig. 3. Decision support system interface

Further, by pressing the *"Calculate"* button, the calculation of informativeness estimation methods is carried out, precisely the Shannon method and the Kullback–Leibler method.

3.2. Data analysis

The experimental study used data on patients suffering from COVID-19 [28]. Figure 4 depicts the histogram of the input data.

Next, we checked the dataset for empty data that would worsen the prediction. Figure 5 shows all data output in terms of data type, presence of zero, and the number of records of 950217 patients.

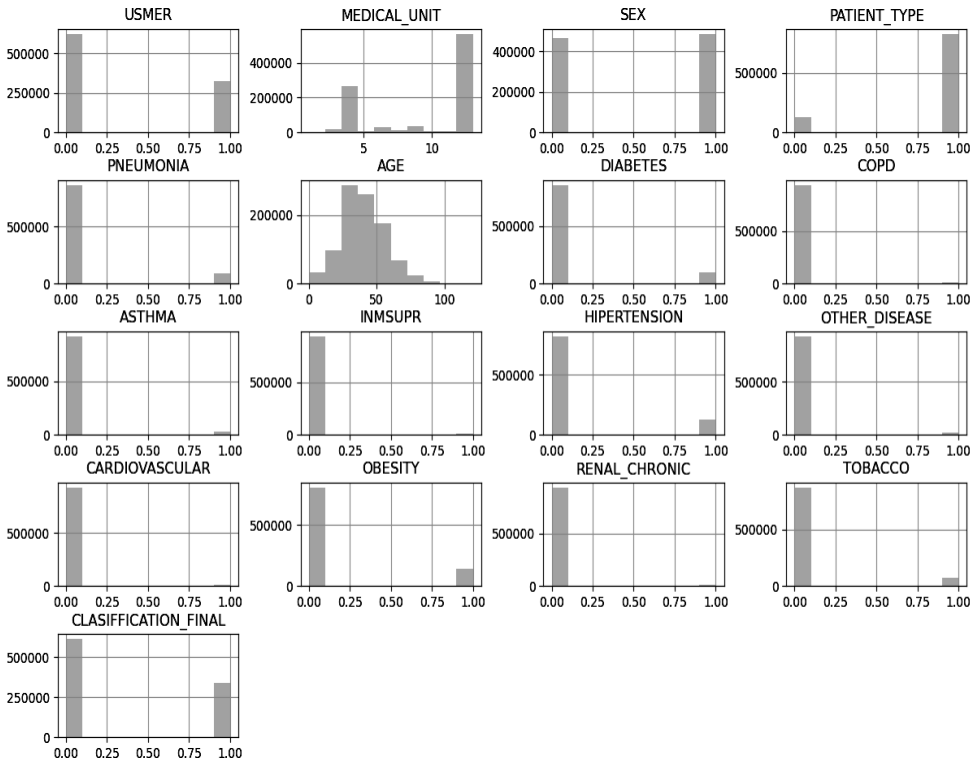
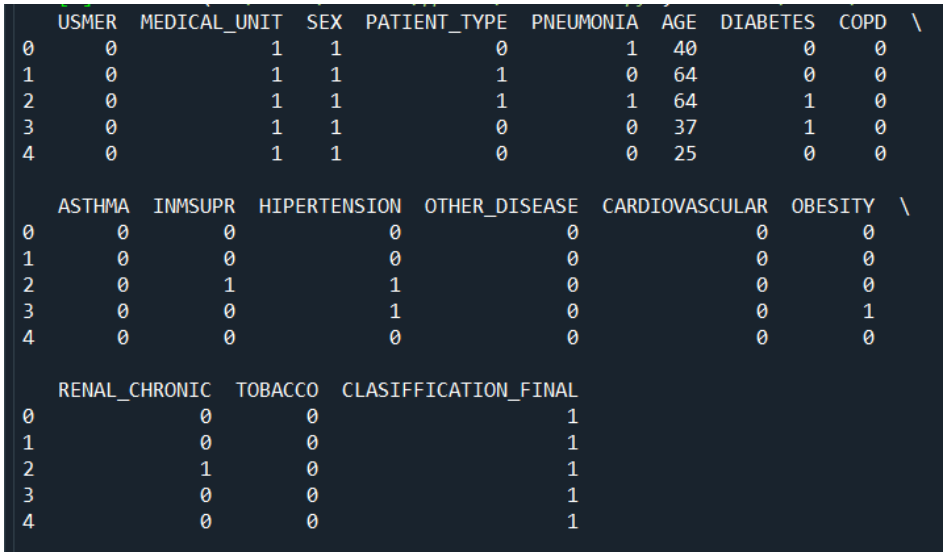


Fig. 4. Patient Data Histogram

```
Data columns (total 17 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   USMER                 950217 non-null  int64
 1   MEDICAL_UNIT          950217 non-null  int64
 2   SEX                   950217 non-null  int64
 3   PATIENT_TYPE          950217 non-null  int64
 4   PNEUMONIA             950217 non-null  int64
 5   AGE                   950217 non-null  int64
 6   DIABETES              950217 non-null  int64
 7   COPD                  950217 non-null  int64
 8   ASTHMA                950217 non-null  int64
 9   INMSUPR               950217 non-null  int64
 10  HIPERTENSION          950217 non-null  int64
 11  OTHER_DISEASE         950217 non-null  int64
 12  CARDIOVASCULAR        950217 non-null  int64
 13  OBESITY               950217 non-null  int64
 14  RENAL_CHRONIC         950217 non-null  int64
 15  TOBACCO               950217 non-null  int64
 16  CLASIFFICATION_FINAL  950217 non-null  int64
dtypes: int64(17)
memory usage: 170.5 MB
```

Fig. 5. Checking for the presence of empty values

Figure 6 shows the output of the first 5 rows of the input data table.



	USMER	MEDICAL_UNIT	SEX	PATIENT_TYPE	PNEUMONIA	AGE	DIABETES	COPD
0	0	1	1	0	1	40	0	0
1	0	1	1	1	0	64	0	0
2	0	1	1	1	1	64	1	0
3	0	1	1	0	0	37	1	0
4	0	1	1	0	0	25	0	0

	ASTHMA	INMSUPR	HIPERTENSION	OTHER_DISEASE	CARDIOVASCULAR	OBESITY
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	1	1	0	0	0
3	0	0	1	0	0	1
4	0	0	0	0	0	0

	RENAL_CHRONIC	TOBACCO	CLASIFFICATION_FINAL
0	0	0	1
1	0	0	1
2	1	0	1
3	0	0	1
4	0	0	1

Fig. 6. View of the first 5 rows of input medical data

3.3. Feature selection

We should note that the Shannon method estimates the informativeness of the investigated recognition in a normalized quantity, which takes values from 0 to 1. Comparison of results of both methods allows the following conclusions: the considered methods do not contradict each other and give similar sets of the most informative features on the same training samples, and the results of the Shannon and Kullback methods mostly coincide. Table shows the results of using methods for assessing the informativeness of features.

Results of calculating the informativeness of features

Name	Results (Shannon)	Results (Kullback–Leibler)
Treatment in medical institutions	0.92	1.55
Medical insurance	0.44	1.99
Gender	0.99	1.73
Patient type	0.55	1.97
Pneumonia	0.43	0.94
Age	0.86	2.00
Diabetes	0.46	0.99
Chronic lung disease	0.08	0.00
Asthma	0.19	0.46
Weakness of the immune system	0.09	0.025
High blood pressure	0.57	1.13
Another disease	0.16	0.34
Cardiovascular disease	0.12	0.18
Obesity	0.60	1.17
Chronic kidney disease	0.10	0.08
Smoking	0.40	0.89
Covid-19 disease	0.93	1.56

The obtained results were visualized. Figures 7 and 8 show which features have an impact and informativeness and which can be excluded from the set.

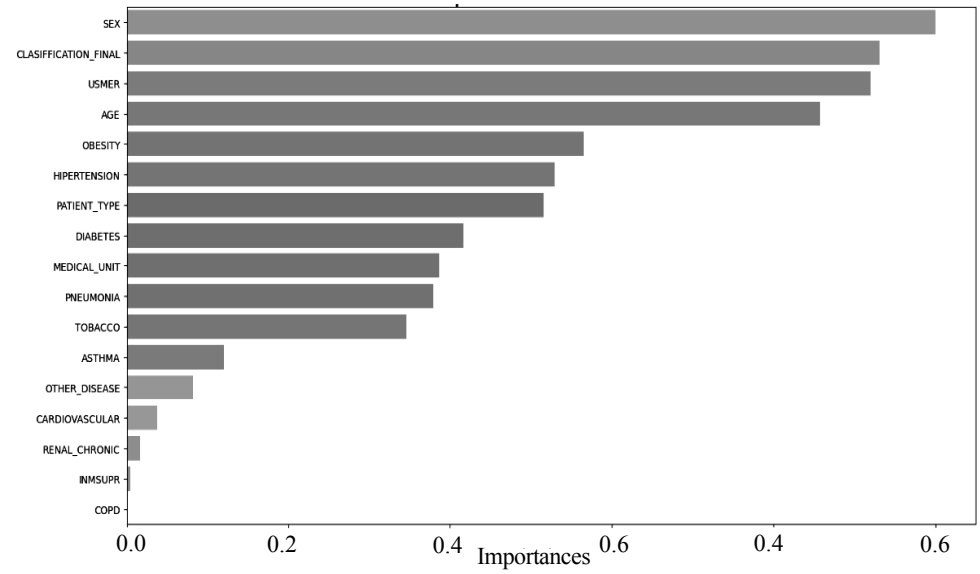


Fig. 7. Diagram of informativeness assessment by the Shannon method

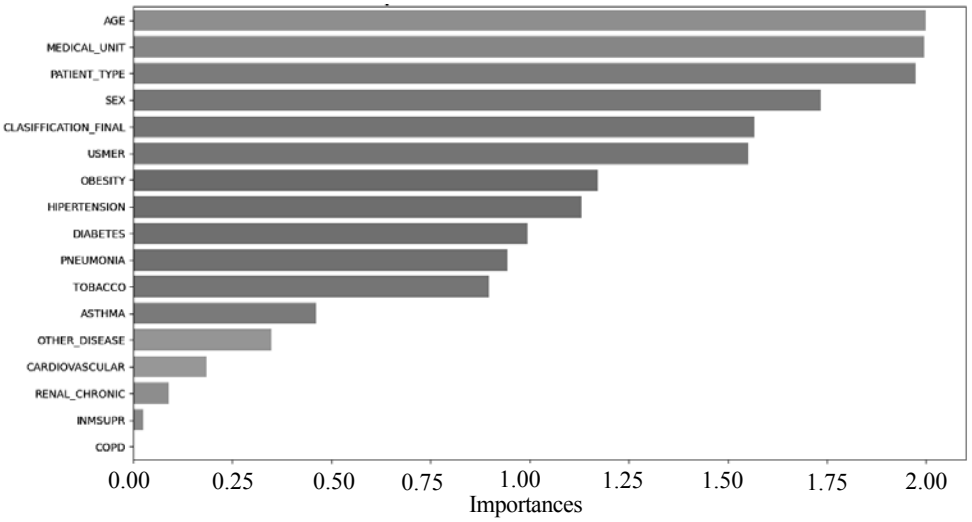


Fig. 8. Diagram of informativeness assessment by the Kullback–Leibler method

4. DISCUSSION

The evaluation of informativeness is pivotal in understanding the dynamics of epidemic processes and devising effective disease control strategies. This study aimed to implement and evaluate methods to assess the informativeness of features that influence epidemic processes. The methods examined in this study, namely the Shannon method and the Kullback–Leibler method, are grounded in the principles of information theory and have distinct advantages, differences, and commonalities. Both methods utilize the concept of event probability and employ a logarithmic scale to measure informativeness, which is particularly helpful when dealing with extremely small or large probability values. These methods are

also extensively applied in machine learning for feature selection, model management, and assessing feature informativeness.

The study found that the Shannon and Kullback–Leibler methods are valuable tools for quantifying the information contained in a random process and thus can be applied across various fields such as information theory, statistics, and machine learning. The comparison of different methods and the results they yield is crucial for understanding their applicability and limitations. It was observed that certain features, such as "Chronic lung disease," "Chronic kidney disease," and "Weakness of the immune system," did not carry significant information for further analysis and prediction, indicating that not all available features are necessarily informative or relevant for epidemic process analysis.

Developing an information system that facilitates the assessment of feature informativeness is a significant contribution of this study. This system not only supports data sample uploading but also enables the calculation of the informativeness of factors that influence the epidemic process. The visualization of the system's results aids in the interpretation and application of the findings.

However, there are several limitations to this study. First, the analysis was based on a specific data set, and the informativeness of features may vary in different contexts or with different diseases. Therefore, the findings of this study may not be directly generalizable to other epidemic processes. Second, the study focused on two specific methods of assessing informativeness, and there may be other methods that could yield different results or insights. Additionally, the study did not consider the potential interactions between different features, which could also influence the informativeness of individual features.

The study contributes a novel perspective by demonstrating a methodical approach to assess the informativeness of various features related to epidemic processes. By applying the Shannon and Kullback–Leibler methods, this study brings a quantitative, data-driven approach to a field often dominated by qualitative assessments and heuristic methods. This quantitative approach can lead to more objective, replicable, and actionable insights into the drivers of epidemic processes.

Additionally, this study contributes by identifying specific features that are not informative in the context of the analyzed data set. This is crucial as it challenges conventional wisdom and prompts a re-evaluation of commonly held beliefs about the most critical factors in driving epidemic processes. This can lead to a paradigm shift in how epidemic processes are analyzed and managed, moving away from a one-size-fits-all approach to a more nuanced, data-driven approach.

Moreover, the study compares two widely used methods for assessing informativeness, thereby providing insights into their relative merits and limitations. This can guide researchers and practitioners in selecting the most appropriate method for their specific context and research questions.

Developing an information system that supports data upload and informativeness calculations adds a practical tool that researchers and practitioners can use to assess the informativeness of features in their own data sets. This contributes to the methodological rigor of future studies and enhances the practical applicability of the findings by enabling real-world implementation.

Future research should validate the findings of this study in different contexts and with different diseases to assess the generalizability of the results. It would also be beneficial to compare the performance of the Shannon and Kullback–Leibler methods with other methods of assessing informativeness. Furthermore, future studies should also explore the potential interactions between different features and their impact on the informativeness of individual features.

Developing and evaluating more sophisticated information systems that can account for feature interactions and other complexities in the data would be a valuable avenue for future research.

Overall, this study contributes a novel perspective, challenges conventional wisdom, provides practical insights into the relative merits of different methods, and offers a practical tool for assessing feature informativeness. These contributions are crucial for enhancing our understanding of epidemic processes and developing more effective strategies for their management.

CONCLUSIONS

The use of methods for assessing informativeness is crucial in analyzing epidemic processes. The main objective of such an analysis is to understand the spread of the disease and determine the effectiveness of strategies to combat it. Methods of informativeness assessment allow for determining how well a specific parameter correlates with the risk of disease. This enables identifying population groups that may be more susceptible to the disease and considering this when developing prevention and treatment strategies.

As a result of this study, methods were identified and implemented that allow assessing the informativeness of features. Methods for assessing the informativeness of features were considered; algorithmic models were developed for the Kullback–Leibler and Shannon methods. Both considered methods are based on information theory principles and have advantages, differences, and standard features. Thus, both the Shannon method and the Kullback–Leibler method are based on the concept of the probability of events, use a logarithmic scale to measure informativeness, which helps in dealing with very small or tremendous probability values, and is widely used in the field of machine learning for evaluating the informativeness of features, model management, and feature selection. Overall, the Shannon and Kullback–Leibler informativeness assessment methods are valuable tools for measuring the information contained in a random process. They can be used in various fields, such as information theory, statistics, machine learning, etc.

Specific examples of using the described algorithmic models are presented. A comparison of different methods and their results was carried out. It was found that such features as “Chronic lung disease”, “Chronic kidney disease”, and “Weakness of the immune system” do not carry information for further work with the table and burden the prediction relative to the presented data set.

An information system for analyzing epidemic process data was developed to assess the informativeness of features. This system supports data sample uploading and calculations of the informativeness of factors affecting the epidemic process. The results of the system operation are visualized.

Acknowledgements. The study was funded by the National Research Foundation of Ukraine in the framework of the research project 2020.02/0404 on the topic “Development of intelligent technologies for assessing the epidemic situation to support decision-making within the population biosafety management”.

REFERENCES

1. K. Batko and A. Ślęzak, “The use of Big Data Analytics in healthcare,” *Big Data*, vol. 9, no. 1 (2022), https://doi.org/10.1186/s40537-021-00553-4.
2. I. Izonin, R. Tkachenko, I. Dronyuk, et al., “Predictive modeling based on small data in clinical medicine: RBF-based additive input-doubling method,” *Mathematical Biosciences and Engineering*, vol. 18, no. 3, pp. 2599–2613 (2021), https://doi.org/10.3934/mbe.2021132.

3. S.Y. Lee, B. Lei, and B. Mallick, "Estimation of COVID-19 spread curves integrating global data and borrowing information," *PLOS ONE*, vol. 15, no. 7, 0236860 (2020), https://doi.org/10.1371/journal.pone.0236860.
4. S. Ma, Y. Sun, and S. Yang, "Using Internet Search Data to Forecast COVID-19 Trends: A Systematic Review," *Analytics*, vol. 1, no. 2, pp. 210–227 (2022), https://doi.org/10.3390/analytics1020014.
5. A. Ibrahim, U. W. Humphries, A. Khan, et al., "COVID-19 Model with High- and Low-Risk Susceptible Population Incorporating the Effect of Vaccines," *Vaccines*, vol. 11, no. 1 (2022), https://doi.org/10.3390/vaccines11010003.
6. N. Davidich, I. Chumachenko, Y. Davidich, et al., "Advanced Traveller Information Systems to Optimizing Freight Driver Route Selection," *2020 13th International Conference on Developments in eSystems Engineering (DeSE)* (2020), https://doi.org/10.1109/dese51703.2020.9450763.
7. S. Fedushko and T. Ustyianovych, "E-Commerce Customers Behavior Research Using Cohort Analysis: A Case Study of COVID-19," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 8, no. 1, pp. 1-12 (2022), https://doi.org/10.3390/joitmc8010012.
8. P.S. Knopov, O.S. Samosonok, and G.D. Bila, "A Model of Infectious Disease Spread with Hidden Carriers," *Cybernetics and Systems Analysis*, vol. 57, no. 4, pp. 647–655 (2021), https://doi.org/10.1007/s10559-021-00390-6.
9. D.A. Klyushin, "Effective algorithms for solving statistical problems posed by COVID-19 pandemic," *Elsevier eBooks*, pp. 21–44 (2023), https://doi.org/10.1016/b978-0-323-90531-2.00005-9.
10. I. Krak, H. Kudin, V. Kasianiuk, et al., "Hyperplane Clustering of the Data in the Vector Space of Features Based on Pseudo Inversion Tools," *CEUR Workshop Proceesings*, vol. 3003, pp. 98–105 (2021), https://ceur-ws.org/Vol-3003/short4.pdf
11. O. Filchakova, D. Dossym, A. Ilyas, et al., "Review of COVID-19 testing and diagnostic methods," *Talanta*, vol. 244, 123409 (2022), https://doi.org/10.1016/ j.talanta.2022.123409.
12. S. Patil, H. Lu, C. L. Saunders, et al., "Public preferences for electronic health data storage, access, and sharing — evidence from a pan-European survey," *Journal of the American Medical Informatics Association*, vol. 23, no. 6, pp. 1096–1106 (2016), https://doi.org/10.1093/jamia/ocw012.
13. V. Berisha, C. Krantsevich, P. R. Hahn, et al., "Digital medicine and the curse of dimensionality," *npj Digital Medicine*, vol. 4, no. 1 (2021) https://doi.org/10.1038/ s41746-021-00521-5.
14. K. Bazilevych, S. Krivtsov, and M. Butkevych, "Intelligent Evaluation of the Informative Features of Cardiac Studies Diagnostic Data using Shannon Method," *CEUR Workshop Proceedings*, vol. 3003, pp. 65–75 (2021).
15. I. Meniailov and H. Padalko, "Application of Multidimensional Scaling Model for Hepatitis C Data Dimensionality Reduction," *CEUR Workshop Proceedings*, vol. 3348, pp. 34–43 (2022).
16. K. O. Bazilevych, D. I. Chumachenko, L. F. Hulianytskyi, et al., "Intelligent Decision-Support System for Epidemiological Diagnostics. I. A Concept of Architecture Design," *Cybernetics and Systems Analysis*, vol. 58, no. 3, pp. 343–353 (2022), https://doi.org/10.1007/s10559-022-00466-x.
17. K.O. Bazilevych, D.I. Chumachenko, L.F. Hulianytskyi, et al., Intelligent Decision-Support System for Epidemiological Diagnostics. II. Information Technologies Development," *Cybernetics and Systems Analysis*, vol. 58, no. 4, pp. 499–509 (2022). https://doi.org/10.1007/s10559-022-00484-9
18. D. Panda, R. Ray, and Satya Ranjan Dash, "Feature Selection: Role in Designing Smart Healthcare Models," *Intelligent systems reference library*, vol. 178, pp. 143–162, (2020), https://doi.org/10.1007/978-3-030-37551-5_9.
19. D. Geiszler, D. A. Polasky, F. Yu, and A. I. Nesvizhskii, "Detecting diagnostic features in MS/MS spectra of post-translationally modified peptides," *Nature Communications*, vol. 14, no. 1 (2023), https://doi.org/10.1038/s41467-023-39828-0.
20. D.E. Ehrmann, S. Joshi, S.D. Goodfellow, et al., "Making machine learning matter to clinicians: model actionability in medical decision-making," *npj Digital Medicine*, vol. 6, no. 1 (2023), https://doi.org/10.1038/s41746-023-00753-7.
21. O. Cliff, M. Prokopenko, and R. Fitch, "Minimising the Kullback–Leibler Divergence for Model Selection in Distributed Nonlinear Systems," *Entropy*, vol. 20, no. 2, p. 51 (2018), doi: https://doi.org/10.3390/e20020051.

22. X. Wang, W. Hou, H. Zhang, et al., "KDE-OCSVM model using Kullback–Leibler divergence to detect anomalies in medical claims," *Expert Systems with Applications*, vol. 200, 117056 (2022), doi: https://doi.org/10.1016/j.eswa.2022.117056.
23. N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, et al., "A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction," *Frontiers in Bioinformatics*, vol. 2 (2022), https://doi.org/10.3389/fbinf.2022.927312.
24. J. Li, K. Cheng, S. Wang, et al., "Feature Selection," *ACM Computing Surveys*, vol. 50, no.6, pp. 1–45 (2018), https://doi.org/10.1145/3136625.
25. F. Jalali-najafabadi, M. Stadler, N. Dand, et al., "Application of information theoretic feature selection and machine learning methods for the development of genetic risk prediction models," *Scientific Reports*, vol. 11, no. 1 (2021), https://doi.org/10.1038/s41598-021-00854-x.
26. A. D. Al-Nasser, A. Rawashdeh, and A. Talal, "On using Shannon entropy measure for formulating new weighted exponential distribution," *Journal of Taibah University for Science*, vol. 16, no. 1, pp. 1035–1047 (2022), https://doi.org/10.1080/ 16583655.2022.2135806.
27. "*Scikit-learn: machine learning in Python,*" *Scikit-learn.org* (2019), https://scikit-learn.org/stable/
28. "COVID-19 Dataset," *www.kaggle.com* (2022), https://www.kaggle.com/datasets/meirnizri/covid19-dataset

Received 06.09.2023

INFORMATION ON THE ARTICLE

Kseniia O. Bazilevych, ORCID: 0000-0001-5332-9545, National Aerospace University "Kharkiv Aviation Institute", Ukraine, e-mail: k.bazilevych@khai.edu

Olena Yu. Kyrylenko, ORCID: 0009-0005-8917-0878, National Aerospace University "Kharkiv Aviation Institute", Ukraine, e-mail: o.kyrylenko@khai.edu

Yurii L. Parfenyuk, ORCID: 0000-0001-5357-1868, V.N. Karazin Kharkiv National University, Ukraine, e-mail: parfuriy.l@gmail.com

Sergiy V. Yakovlev, ORCID: 0000-0003-1707-843X, National Aerospace University "Kharkiv Aviation Institute", Ukraine, e-mail: s.yakovlev@khai.edu

Serhii O. Krivtsov, ORCID: 0000-0001-5214-0927, National Aerospace University "Kharkiv Aviation Institute", Ukraine, e-mail: krivtsovpro@gmail.com

Ievgen S. Meniailov, ORCID: 0000-0002-9440-8378, V.N. Karazin Kharkiv National University, Ukraine, e-mail: evgenii.menyailov@gmail.com

Victoriya O. Kuznietcova, ORCID: 0000-0003-3882-1333, V.N. Karazin Kharkiv National University, Ukraine, e-mail: vkuznietcova@karazin.ua

Dmytro I. Chumachenko, ORCID: 0000-0003-2623-3294, National Aerospace University "Kharkiv Aviation Institute", Ukraine, e-mail: d.chumachenko@khai.edu

ІНФОРМАЦІЙНА СИСТЕМА ДЛЯ ОЦІНЮВАННЯ ІНФОРМАТИВНОСТІ ОЗНАК ЕПІДЕМІЧНОГО ПРОЦЕСУ / К.О. Базілевич, О.Ю. Кіріленко, Ю.Л. Парфенюк, С.В. Яковлев, С.О. Кривцов, Є.С. Меняйлов, В.О. Кузнецова, Д.І. Чумаченко

Анотація. Роботп полягає в оцінюванні інформативності параметрів, які впливають на епідемічні процеси, з використанням методів Шенона та Кульбака–Лейблера на основі їх фундаментальності у принципах теорії інформації та їх широкого застосування в машинному навчанні, статистиці та інших відповідних галузях. Проведено порівняльний аналіз результатів, отриманих обома методами, розроблено інформаційну систему для спрощення завантаження вибірок даних та обчислення інформативності факторів, які впливають на епідемічні процеси. Показано, що деякі ознаки, такі як «хронічне захворювання легень», «хронічне захворювання нирок» та «ослаблений імунітет», не містили значущої інформації для подальшого аналізу та ускладнювали процес прогнозування за даними досліджуваного набору даних. Розроблена інформаційна система ефективно підтримує оцінювання інформативності ознак, тим самим сприяючи комплексному аналізу епідемічних процесів, візуалізації результатів, а також поточному стану знань. Надано конкретні приклади застосування описаних алгоритмічних моделей, порівняння різних методів та їх результатів та розроблення підтримувального інструменту для аналізу епідемічних процесів.

Ключові слова: інформаційна система, епідемічний процес, інформативність ознаки, метод Шенона, метод Кульбака–Лейблера.