MEANS FOR MEASURING THE THERMAL QUANTITIES

MACHINE LEARNING METHODS IN THERMOMETERS' DATA EXTRACTION AND PROCESSING

Pylyp Skoropad, Dr. Sc., Prof.; Andrii Yuras, PhD Student
Lviv Polytechnic National University, Ukraine; e-mail: andrii.o.yuras@lpnu.ua

https://doi.org/10.23939/istcmtm2024.02.040

Abstract. Research focuses on developing an all-encompassing algorithm for efficiently extracting, processing, and analyzing data about thermometers. The examination involves the application of a branch of artificial intelligence, in particular machine learning (ML) methods, as a means of automating processes. Such methods facilitate the identification and aggregation of pertinent data, the detection of gaps, and the conversion of unstructured text into an easily analyzable structured format. The paper details the employment of reinforcement learning for the automatic extraction of information from diverse resources, natural language processing for analysis of textual values, and the decision tree method for discerning patterns within the data.

Key words: Artificial intelligence, machine learning, prediction, temperature measurement, thermometer.

1. Introduction

Temperature measurement is an essential component in science, medicine, industry, etc. Thermometers, as the primary means of measurement, differ by various parameters, such as accuracy, measurement range, and inertia. Despite available information about these means, searching, extracting, and processing can be time-consuming.

2. Drawbacks

Parameter Selection: Adjusting the settings of machine learning models, such as the learning rate and discount factor in the Q-learning algorithm, can pose a challenge. Incorrectly chosen parameters can lead to poor results.

Gap Detection: The algorithm can identify gaps, but it does not always mean that they can be filled.

Data Constraints: The results depend on availability and quality. If the data is incomplete, inaccurate, or biased, it can impact the quality of the model and its predictions.

Reality Correspondence: Since the model is based on data from the Internet, the results may not correspond to reality. Information about thermometers can change over time, and these changes may not be reflected in the model.

3. Goal

Develop an algorithm based on machine learning methods for searching, extracting, and processing data about measurement devices.

4. Investigation of data accumulation and processing methods

Data extraction is the process of obtaining various types of values from diverse sources, many of which may be poorly organized or completely unstructured. Extraction facilitates the consolidation, processing, and refinement of data so that it can be stored in a centralized location for transformation. Locations can be local, cloud-based, or hybrid.

Web scraping is the process of using bots to obtain content and information from a website, widely employed for various purposes, including online price comparison, monitoring changes in weather data, detecting changes on a website, integrating from multiple sources, receiving offers and discounts, collecting job postings from job portals, brand monitoring, and market analysis [1]. It serves as a tool for quick and efficient data selection. Research on web citation analysis aids in discovering pertinent articles for further exploration [2]. The investigation involves an analysis of a methodology grounded in the Firefly optimization algorithm. It is being designed to obtain web citation-related values based on a specific query. In a similar vein, [3] employs web scraping based on the constructed HTML DOM algorithm. They utilize this to summarize the publication of scientific articles from Google Scholar. These summaries are formatted for presentation as a report in PDF or Excel format. In [4], a special system for searching news on the Internet is presented. Mechanism focuses on creating a news repository by extracting content from a network information sheet, obtained from various electronic information portals.

API or Application Programming Interface, is a set of rules and protocols for programs to interact with each other. In the context of processing, the API is used to extract data from various sources. In [6], the focus is on the exploration of social networks for analysis and mining purposes. The aim is to study specific groups (political or regional), detect these groups, and create strategic plans to combat terrorism and crime, with the aid of social media information and networks. In [7], an API is developed for creating an archive from networks. The impact of various APIs on data access and collection creation is examined.

Data processing includes - cleaning: removal of errors, incorrect values, duplicates, and gaps; transformations: changing the format or structure, to make them convenient for analysis (conversion of textual to numerical); normalization: bringing to a common scale so they can be compared or analyzed; sampling: selection of a subset for analysis; outlier detection: detection and handling values that significantly differ from other values in the set [13].

Drawing on their research, software was developed to perform the task of accumulating information about thermometers. An example of web scraping grounded on the Firefly optimization algorithm was employed to extract data from websites [2], enriched sampling with HTML DOM [3] to extract detailed values about devices, and processed them following the method of organizing a news repository for data aggregation [4]. To understand how it works in the study, the principle of its operation is presented in the form of a flowchart in Fig.1. To the API, the approach described in the study [6] for mining was adapted, as well as the methodology for the creation of a value collection [7]. Strategy is visualized as a flowchart in Fig.2. The outcomes of the described methods are presented in Table 1. It is important to note that the process of manual source search proved to be time-consuming. It is attributed to the dispersed nature of information about temperature measuring devices across various websites, often lacking appropriate structure or systematization. Some websites might lack sufficient information or even present inaccurate data, significantly complicating the search and extraction process. With several potential sources, the manual procedure can be quite extensive. It encompasses not just the search but the verification of information accuracy, along with its analysis and systematization. Post extraction in both algorithms, a processing phase ensues, which guarantees the quality of data and the precision of subsequent analysis. Data processing requires deep knowledge in the field from which this data was obtained. In addition, it is necessary to understand which parameters are important for a specific task. This includes understanding the importance of attributes, determining which ones can be removed, and detecting the data that needs to be transformed for further analysis.

From Fig.1, the working principle of the data extraction algorithm can be illustrated as follows: a) Defining the scraping goal: The data to be collected is determined, for instance, temperature data from a specific site; b) Selecting web pages for scraping: Pages are selected from which data will be gathered; c) Development of an extraction script: Code is created that navigates to web pages, locates the necessary data, and accumulates them; d) Script execution: The script is run to collect data; e) Saving the gathered data for future utilization; f) Consistency check: Conducting checks for errors or omissions; g) Removal of unnecessary attributes: Data not essential for analysis is discarded; h) Conversion to the required format Removal of unnecessary attributes (from text to numeric); i) Outlier analysis: Checking for the presence of outliers that may influence the analysis results; j) Selection of relevant data for analysis: Filtering is done to retain only the data necessary for analysis.

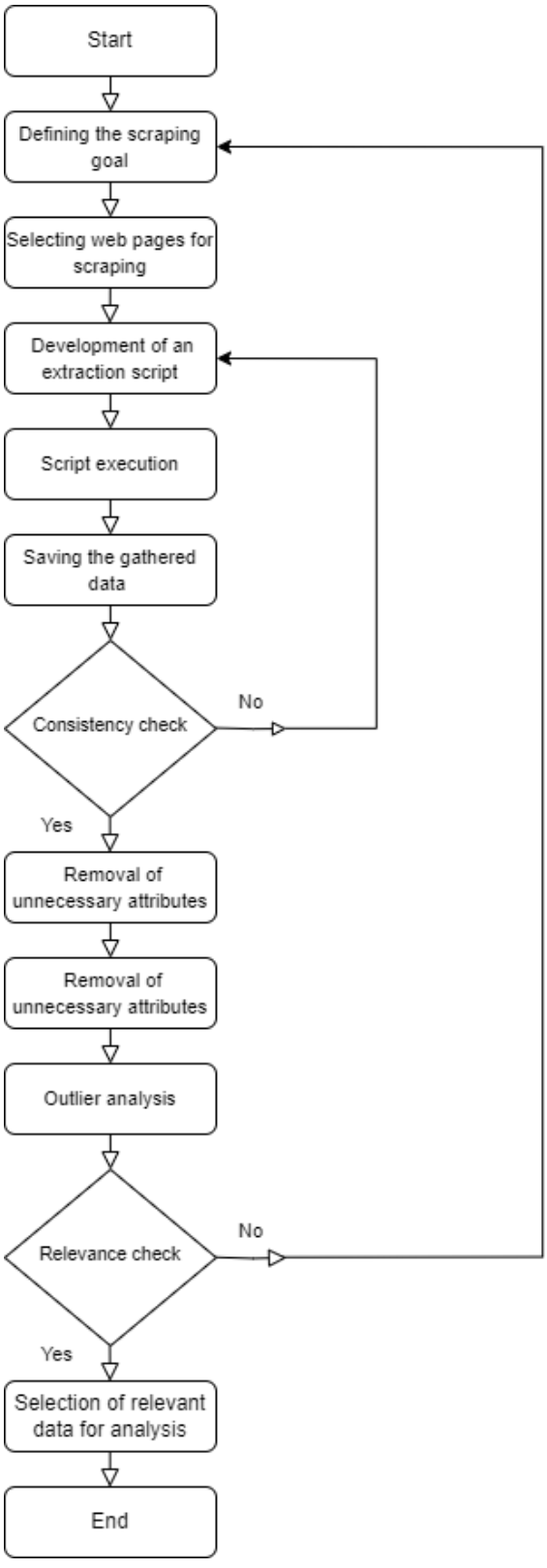

Fig.1. Web Scraping Data Extraction Algorithm

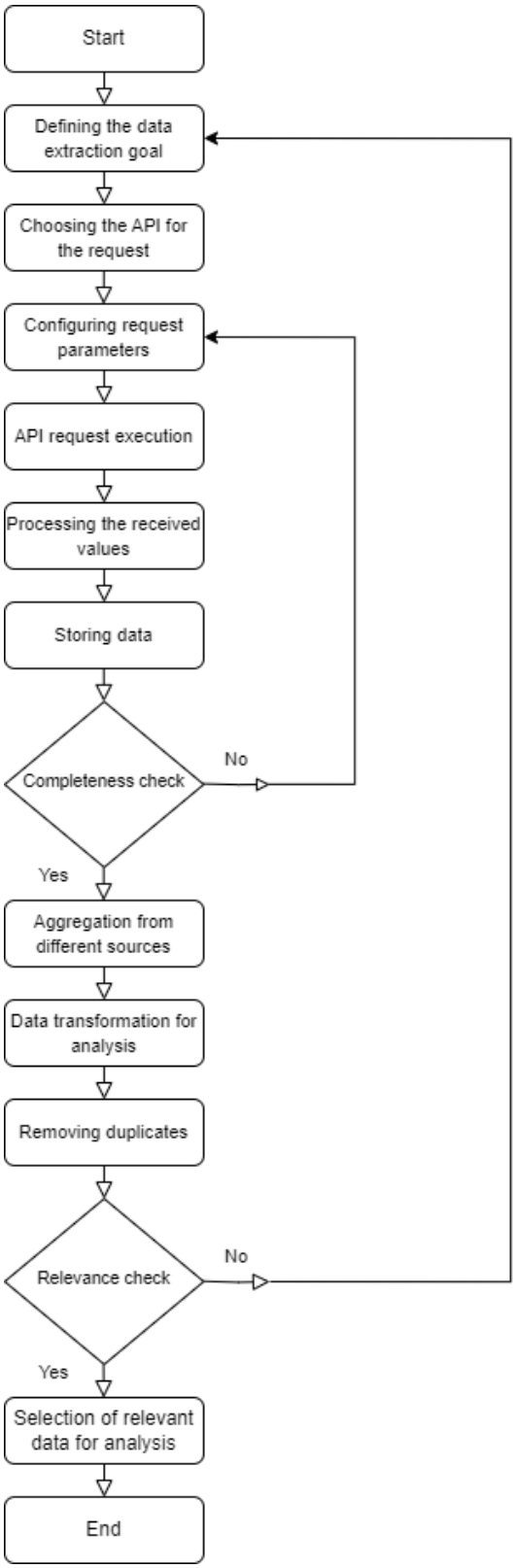

Fig.2. Data extraction algorithm with API requests

Fig.2 illustrates the data acquisition algorithm involving API requests, which operates as follows: a) Defining the data extraction goal: The data to be collected is defined. For example, temperature data from a specific API; b) Choosing the API for the request: The API for information collection is selected; c) Configuring request parameters: The request parameters, such as type, time interval, etc., are set; d) Performing the request: The request is sent to the API, and the results are received; e) Processing the received values: These are converted into the required format; f) Storing: The data is stored for further utilization; g) Completeness check: It is verified whether all necessary data has been received and there are no omissions; h) Aggregation from different sources: If the information has been collected from different APIs, it can be combined into one set; i) Data transformation for analysis: The data is converted into a format suitable for analysis; j) Removing duplicates: If there are duplicates, they are removed; k) Selection of relevant data for analysis: Filtering is done to retain only the data necessary for analysis.

In the context of gathering information about temperature measuring devices, existing technical approaches are not suitable. This is due to their responsibility only for data selection. These approaches only select from predefined sources, not providing processing or interpretation. Consequently, an all-encompassing algorithm employing ML methods was devised, targeting the enhancement of data extraction process efficiency. It includes the automation of detecting relevant information and its processing. The result of the work is shown in the form of a flowchart in Fig.3. In the study, we refer to the ideas described by Li [6], who applies natural language processing (NLP) to reduce the time required for gap detection in extraction results. By ranking phrases based on the descending frequency range, a set of keywords that are absent in the results is obtained, and the research gap is determined. Similar was described by Clément [8] in his work. The developed algorithm is expanded by adding ML models at different stages.

To accumulate and aggregate data, reinforcement learning was applied [9-11]. The basis of this algorithm is the Q-function [14], which represents the expected total reward for a state-action pair. The value of this function is updated according to the formula:

$$Q(s,a) = Q(s,a) + \alpha * (r + \gamma * \max(Q(s',a')) - Q(s,a)), \quad (1)$$

where Q(s, a) is the value of the Q-function for the state s (current source) and action a (extraction of information about each type of thermometer), α is the learning rate, r is the reward (successful extraction of relevant values), γ is the discount factor, which controls the significance of future rewards, max Q(s', a') is the highest value of the Q-function for the next state s' (new resource for extraction) and all possible actions a' (obtaining different types of thermometers). When choosing an action, the ε-greedy strategy [15] was applied. According to this strategy, the model with a probability of (1 - ε) chooses the action that

maximizes Q(s, a), and with a probability of ε chooses an action randomly. Balancing the exploration and exploitation process prevented the model from getting stuck in a local optimum. The ML method enabled the identification and collection of relevant data from various resources. The model "learned" its mistakes and successes, constantly adjusting actions to achieve better results. As a result, the model can perform extraction without the need for manual intervention, increasing the efficiency of the process. The NLP method works to analyze text values related to thermometers. This process begins with tokenization, where the input text S is converted into a list of separate elements T = [t1, t2, ..., tn], where ti is the i-th token in the text. The next step is removing stop words. For this, a list of words S is created that does not carry specific information about thermometers, and all such tokens are removed from the list T, obtaining a new list of tokens T' = [ti | ti is in T, ti, not S]. A step further is stemming and lemmatization. Transforming each token ti into its "stem" (base) or "lemma" (normal form), obtaining the final list of tokens T" = [stem(ti) or lemma(ti) | ti is in T']. The facilitation of detecting gaps in the data and converting unstructured text into a structured format for easy analysis was achieved.

For pattern identification and analysis, the decision tree methodology is deployed. Conceptually, it can be imagined as a data structure where each internal node is responsible for a test on a certain feature, each branch corresponds to the result of this test and each leaf represents the predicted value of the target variable [16]. Mathematically, it was constructed with the ID3 algorithm [12]. This method divides the data into subsets based on different features of thermometers. The main criterion for choosing a feature for splitting is information gain, which measures a reduction in entropy - a measure of uncertainty or "chaos" [16]. Mathematically, entropy is calculated by the well-known formula:

$$(S) = -\sum^{n} \quad p_i log_2(p_i) \ , \quad (2)$$

where S is the sample of thermometers, and p is the number of thermometers with a certain feature. The information gain for each thermometer feature A is calculated by:

$$Gai(A) = H(S) - \sum \quad \frac{|Sv|}{|S|} * H(Sv), \quad (3)$$

where Sv is a subset of the sample S where feature A has the value v. The algorithm recognizes which thermometer features affect the target variable (inertia, measurement range, accuracy) and accordingly make adjustments to achieve the necessary prediction accuracy.

Similar to the previous two algorithms, to understand the working principle of the developed one, it is shown in Fig. 3.

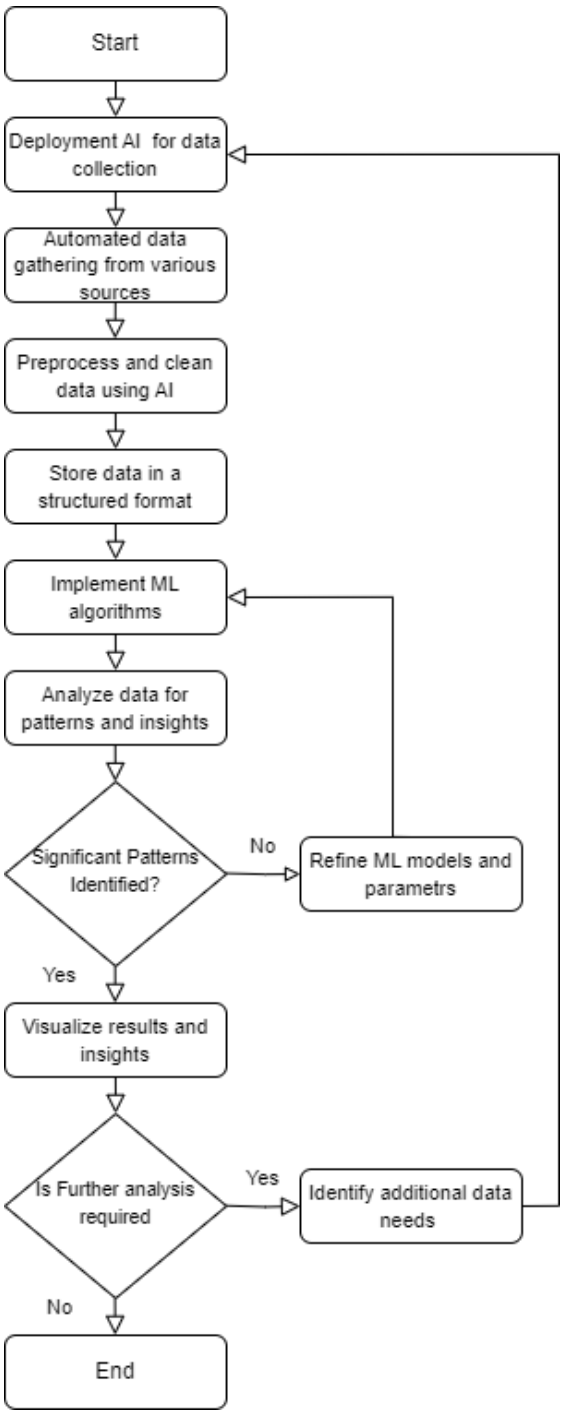

Fig.3. Algorithm for employing ML methods for data extraction and processing

Fig.3 illustrates the operation of the developed algorithm for information extraction utilizing ML methods. It functions as follows: a) Implementation of the developed ML algorithm for identifying and initiating aggregation of relevant values from various resources; b) Automated collection from various sources: The algorithm systematically collects information on various platforms and

repositories without manual intervention; c) Preprocessing and cleaning with ML: This includes cleaning, normalization, and formatting to ensure their readiness for use; d) Storing cleaned data in a structured format: Accumulating the data in a structured format conducive to further processing; e) Running ML algorithms: Designed to analyze organized information, find patterns, and correlations; f) Analyzing for patterns: Detecting significant patterns and deriving practical insights; g) Visualization of results: Employing visualization tools to present identified patterns and understand the overall situation in a comprehensive form; h) Refinement cycle: If the initial analysis doesn't reveal significant patterns, there's an iterative cycle to improve ML models and their parameters to refine results; i) Decision on further analysis: Based on the completeness and adequacy of the results, a decision is made on the need for further review, which could potentially lead to the identification of needs for additional values.

To demonstrate the effectiveness of the developed algorithm, the results of its work are shown in Table below.

It's important to note that the algorithm with ML models has increased the efficiency of the extraction and processing process. It has also reduced the amount of time needed for tasks, demonstrating the ability to independently search for relevant sources about thermometers. The determination of which sources contained the necessary information was made upon analyzing them, resulting in a reduction of resources that required manual verification.

Table. The effectiveness of the developed algorithm

Characteristic	Web Scraping	API-based Data Collection	ML-based Data Collection
Number of sources	25 websites	7 API	67 sources
Volume of data obtained (MB)	155	104	429
Number of gaps/errors	316	268	41
Number of unnecessary attributes	11	8	2
Number of unique values	55.2%	42.4%	71%
Number of relevant values for analysis (MB)	52	38	189
Number of successful queries	204	526	2819
Percentage of query deviations	24%	5%	17%
Number of detected patterns (for ML)	N/A	N/A	5

Compared to web scraping and data collection via API, the ML model-based algorithm proved to be the most effective. It processed the largest number of sources, 57, while providing the largest volume of data (429 MB). Despite the large number of processed sources, this method ensures the fewest errors and gaps, 41, indicating its accuracy. It also found the fewest unnecessary attributes, 2, indicating its effectiveness in data filtration. In terms of uniqueness and relevance of data, it outperformed others: it identified 71% unique values and provided 189 MB of relevant data for analysis. In addition, it identified 5 patterns, which is impossible using other methods. Thus, the developed algorithm is the best option for collecting and processing data about thermometer parameters. In addition, there are advantages that are difficult to quantify numerically. This includes automation, adaptability, scalability, and continuous improvement.

5. Conclusions

Current study has explored traditional data mining methods like Web Scraping and API in searching for information about measuring devices. A comprehensive algorithm, drawing on ML methods for efficient data extraction and processing, was developed. Its effectiveness significantly surpasses traditional methods in data extraction and analysis efficiency. The findings suggest that integrating ML methods into the data collection and processing stage can enhance the process's efficiency. This could potentially result in a productivity surge of 15% or more, emphasizing the importance of ML algorithms when opting for optimal equipment.

6. Gratitude

The authors are grateful to the Staff of the Department of Information-Measuring Technologies for their mild assistance in designing the article for publication.

7. Mutual claims of authors

The authors declare the absence of any financial or other potential conflict related to the work.

References

[1] Nguyen, V. H., Sinnappan, S., and Huynh, M. (2021). Analyzing Australian SME Instagram Engagement via Web Scraping. Pacific Asia Journal of the Association for Information Systems, 13(2):11-43. Available: https://aisel.aisnet.org/pajais/vol13/iss2/2/

[2] Seliverstov, Y., Seliverstov, S., Malygin, I., and Korolev, O. (2020). Traffic safety evaluation in Northwestern Federal District using sentiment anal- ysis of Internet users' reviews. Transportation Re- search Procedia, 50:626-635. Available: https://doi.org/10.1016/j.trpro.2020.10.074

[3] E. Suganya, S. Vijayarani, "Firefly Optimization Al- gorithm Based Web Scraping for Web Citation Ex- traction," Wireless Personal Communications, vol. 118, no. 2, May 2021. DOI:10.1007/s11277-021-08093-z.

[4] Rahmatulloh, A., and Gunawan, R. (2020). Web Scraping with HTML DOM Method for Data Collec- tion of Scientific Articles from Google Scholar. In- donesian Journal of Information Systems, 2(2):95-104. DOI:10.24002/ijis.v2i2.3029

[5] S. Kolli, P. Rama Krishna, P. Balakesava Reddy, "A Novel NLP and Machine Learning Based Text Ex- traction Approach from Online News Feed," May 2021. [Online]. Available: https://www.re- searchgate.net/publica- tion/351902660_A_NOVEL_NLP_AND_MACHINE_LEARNING_BASED_TEXT_EXTRACTION _APPROACH_FROM_ONLINE_NEWS_FEED.

[6] Li, R. Y. M. (2020). Building updated research agenda by investigating papers indexed on Google Scholar: A natural language processing approach. In International Conference on Applied Human Factors and Ergonomics. Springer, Cham:298-305. DOI:10.1007/978-3-030-51328-3_42

[7] Nicolas, C., Kim, J., and Chi, S. (2021). Natural lan- guage processing-based characterization of top-down communication in smart cities for enhancing citizen alignment. Sustainable Cities and Society, 66:102674. Available: https://doi.org/10.1016/j.scs.2020.102674

[8] Zhou, N. Duan, S. Liu, H.-Y. Shum, "Progress in Neural NLP: Modeling, Learning, and Reasoning," Engineering,[Online]. Available: https://doi.org/10.1016/j.eng.2019.12.014

[9] O. Lopatko, I. Mykytin, "Neural networks as a means of predicting the temperature value during the transient process," Measuring Equipment and Metrology: Inter- departmental Scientific and Technical Collection, vol. 77, pp. 65-70, 2016. Available: http://www.irbis- nbuv.gov.ua/cgi-bin/irbis_nbuv/cgiirbis_64.exe?I21DBN=LINK&P21DBN=UJRN&Z21ID=&S21RE F=10&S21CNR=20&S21STN=1&S21FMT=ASP_meta&C21COM=S&2_S21P03=FILA=&2_S21STR=metrolog_2016_77_11

[10] O. Lopatko, I. Mykytyn, "Predicting the temperature of water and air flows using a neural network," Measuring Equipment and Metrology: Interdepart- mental Scientific and Technical Collection, vol. 79, no. 3, pp. 37-41, 2018. Available: https://journals. indexcopernicus.com/search/article?articleId= 2064465

[11] O. Lopatko, I. Mykytyn, "Predicting the temperature value using neural networks," All-Ukrainian Scien- tific and Practical Conference "Industrial Automa- tion in Ukraine. Education and Training," Lviv, 2016, pp. 57-58. Available: https://lpnu.ua/sites/de- fault/files/2020/dissertation/1498/areflopatkooo.pdf

[12] Z. Liu, X. Pan, "Comparison and analysis of applica- tions of ID3, CART decision tree models and neural network model in medical diagnosis and prognosis evaluation," Journal of Clinical Images and Medical Case Reports, vol. 2, 2021. DOI:10.52768/2766- 7820/1101

[13] K. Maharana, S. Mondal, B. Nemade, "A review: Data pre-processing and data augmentation techniques," [Online]. Available: https://doi.org/10.1016/j.gltp. 2022.04.020.

[14] I. A. Zamfirache, R.-E. Precup, R.-C. Roman, E. M. Petriu, "Reinforcement Learning-based control using Q-learning and gravitational search algorithm with experimental validation on a nonlinear servo system" DOI:10.1016/j.ins.2021.10.070

[15] C. Dann, Y. Mansour, M. Mohri, A. Sekhari, K. Sri- dharan, "Guarantees for Epsilon-Greedy Reinforcement Learning with Function Approximation," in Proceedings of the 39th International Conference on Machine Learning, PMLR, vol. 162, pp. 4666-4689, 2022. Available: https://doi.org/10.48550/arXiv.2206.09421

[16] G. Singer, I. Cohen, "An Objective-Based En- tropy Approach for Interpretable Decision Tree Mod- els in Support of Human Resource Management: The Case of Absenteeism at Work," Faculty of Engineer- ing, Bar-Ilan University, Ramat-Gan 52900, Israel. DOI: 10.3390/e22080821