TEM Journal. Volume 13, Issue 1, pages 797-808, ISSN 2217-8309, DOI: 10.18421/TEM131-82, February 2024.

# Unlocking Automated Machine Learning Efficiency: Meta-Learning Dynamics in Social Sciences for Education and Business Data

Dijana Oreški [1], Dunja Višnjić [1], Nikola Kadoić [1]

[1] *University of Zagreb Faculty of Organization and Informatics, Pavlinska 2, Varaždin*

***Abstract*** **– Automated Machine Learning (AutoML) utilizing meta-learning (M-L) has gained prominence in the scientific community. Current M-L methods necessitate substantial data and computational resources for extracting meta-features encoding data properties. However, the time needed for meta-feature extraction exceeds that for predictions in M-L systems. This article proposes a domain-specific M-L paradigm tailored to social science, aiming to identify universally applicable meta-features in social science data. Investigating domain-specific properties, the study discerned common meta-features across social science domains, facilitating an efficient AutoML strategy with reduced data requirements. Ninety meta-features, clustered into eight groups characterizing social science data, were employed, focusing on education and business domains. An analysis of 46 datasets revealed domain-specific variations in meta-feature values, confirmed by Wilcoxon tests. Notably, certain meta-features exhibited consistency across social science domains, demonstrating potential for cross-domain AutoML adoption. This research introduces a targeted M-L approach, optimizing AutoML efficiency for social science applications by identifying common meta-features across diverse domains.**

***Keywords*** **– Meta learning, meta-features, domain meta-learning, domain-specific machine learning.**

DOI: 10.18421/TEM131-82
https://doi.org/10.18421/TEM131-82

**Corresponding author:** Dijana Oreški,
*University of Zagreb Faculty of Organization and Informatics, Pavlinska 2, Varaždin*
**Email:** dijana.oreski@foi.hr

*Received: 03 October 2023.*
*Revised: 22 December 2023.*
*Accepted: 16 February 2024.*
*Published: 27 February 2024.*



## 1. Introduction

Many machine learning algorithms (MLA) are developed and applied in a wide spectrum of domains. Their application can be time-consuming and complex and it is necessary to streamline the procedure of choosing algorithms automatically. The "no free lunch" theory states that there is not one best algorithm that works in every circumstance [43], [28]. Current approaches rely on "trial and error," which is inadequate for solving complicated issues. The meta-learning (M-L) was developed by applying a data-driven methodology and learning from experience. In the realm of MLA, M-L refers to the approach of acquiring knowledge from past experiences, gained through the application of diverse algorithms on various datasets [11]. This concept encompasses techniques that can incorporate information about both datasets and models (such as configuration and performance metrics). The data used in M-L approaches is called meta-data, where meta-features (MF) are extracted so M-L can be performed. Extracting MF is the initial step in the M-L process, which is a challenging task. The meta-feature can be regarded as a collection of metrics designed to consistently depict the attributes of distinct problems [16].

The success of M-L relies significantly on the type and quality of MF [9]. Therefore, it is vital to examine a diverse array of potential candidates [57].

In the existing literature, researchers have put forth different sets of MF, with these features being notably contingent on the nature of the problem at hand [58]. Consequently, it becomes essential to identify suitable MF tailored to specific problem types [16]. However, prior investigations into MF have not specifically targeted particular domains.

By applying the “no free lunch” theory in this context, it is possible to conclude that MF values for a dataset in one domain may not be the same in another domain. Thus, this paper deals with three research questions:

a. What are the properties of education datasets?
b. What are the properties of business datasets?

c. Do the properties differ depending on the domain?

This study is motivated by two primary objectives. First, the data properties of the education and business datasets were empirically examined. Second, by using a statistical test, a comprehensive comparison of the properties of these different domains, measured by MF, was performed. Hence, we are directly tackling the research requirements highlighted by:

a. Chu *et al.* [16], MF are problem-dependent, and more studies should identify appropriate MF tailored to specific problem categories.
b. Monteiro *et al*., the principal source of challenges in MLA stems from the novel properties exhibited by data [41]. Furthermore, the M-L approach is particularly important for business domains, since it requires a fast deployment of analytical techniques.
c. Romero *et al.*, studies addressing MF are limited in the number of MF that are used [47]. Their conclusion emphasizes the necessity for a more extensive range of MF to effectively capture the relevant features of educational data.
d. Kanda *et al*., the success of M-L depends on the quality of the MF [29].

The following sections of the paper are structured as follows: Section 2 provides the base regarding M-L and data properties measured by MF. In section 3, the research methodology is outlined, encompassing 8 groups of MF, 46 datasets derived from 2 domains within the social sciences, and the utilization of the Wilcoxon rank sum test for discerning distinctions between two independent sample groups. Section 4 showcases the meta-feature performance across all datasets, examining variations between business and education data. Finally, section 5 concludes the paper by presenting guidelines for future research.

## 2. Background

The application of M-L for algorithm selection involves employing the automated machine learning (AutoML) approach. This approach aims to generate meta-knowledge by establishing connections between data properties, represented by MF, and subsequently assessing the performance of MLA.

Researchers have explored the application of M-L to address challenges in algorithm selection [41]. Various studies, such as [17] and [25], suggest the utilization of MF to enhance the AutoML process. MF provides insights into the correlations between data properties and the efficacy of MLA, offering a basis for selecting the most suitable algorithm for a novel problem [54]. The use of M-L to select algorithms has been studied on the general level. Several studies have made significant contributions in the past, projects such as STATLOG [32], METAL [12], and NOEMON [3]. Several papers highlight that the features of a dataset play a crucial role in influencing the performance of MLA, which demonstrates that the dataset's MF can determine which algorithm is optimal [15], [20], [36], [31]. Ali and Smith's study prove that understanding the dataset properties is required for a learning algorithm selection, whereas Song and Wang pointed out challenges of the optimal MLA selection since it depends on the dataset that is being used [53]. Zhang *et al.* concentrated on the characteristics of one dataset, having attempted to determine which technique was better by thoroughly comparing various approaches [61]. Bogatinovski *et al.* conducted a comprehensive meta-learning study to date, where 40 datasets together with 50 MF were analysed [10]. Using meta-modeling a correlation between MF and technique accuracy was found. Monteiro *et al.* argue that the increase in dataset complexity makes it more challenging for an expert to comprehend the MF and therefore to select the optimal algorithm [41]. Lorena *et al*. emphasized the need for a data-driven approach for an efficient algorithm selection method, as well as the significance of investigating datasets' domain properties. The reasoning for this is the assumption that similar datasets/domains should have similar learning patterns when applying MLA [39].

Wu and Lu claim that although some researchers have contributed to the automation of algorithm selection based on data features in some domains, this does not apply to datasets in other domains [60]. Only a few papers indicate domain-specific M-L. As an illustration, Sivakumar *et al.* conducted a comparison of algorithm performance in the medical domain, specifically focusing on early cancer diagnosis, where classification methodology, based on their examination of supervised learning algorithms on various datasets, was proposed [52]. Garouani *et al*. investigated the manufacturing domain by creating AMLBID (AutoML tool for Big Industrial Data), which is a novel AutoML system that relies on M-L, which generates a ranked list of all candidate algorithms given a dataset, based on their expected performance and the desired evaluation metric (e.g., predictive accuracy, precision, or recall) [27].

This paper makes a noteworthy contribution to domain-specific M-L, particularly emphasizing the domain of social sciences. The pervasive influence of digital systems on our daily activities results in substantial data generation during user interactions. This data, utilized in social science research, is characterized by its complexity and dynamism, encompassing both technological elements and social interactions.

Two different areas of social science are examined in this study, business, and education. Academic behavior and achievement are predicted based on student data that is collected from kindergarten through higher education.

Li, Wang, and Wang underscored the importance of taking into account data properties during the development of predictive models in the field of education [38], while Cui *et al*. emphasized the need to improve MLA applications and provide insight into the performance of algorithms for specific problems [18].

The business domain is indirectly influenced by education, since upon graduation, students enter the business world. According to Bergmann *et al.*, the entrepreneurial domain, with its specificities, lacks a systematic study of relationships between dataset properties and methodological capabilities [7]. Researchers that study entrepreneurial activity, for instance, should be aware that they are dealing with "rare events" (class imbalance problem, where one value of the dependent variable occurs more frequently than the other).

Literature review showed that intelligent data analysis is insufficiently represented in social sciences, not having any research that is focused on examining the characteristics of datasets specifically in social sciences. Also, there are no M-L frameworks for specific social science problems. This research intends to fill these gaps.

## 3. Research Design

The proposed approach comprises three steps. Initially, datasets were extracted from publicly available repositories. Subsequently, meta-feature values were calculated for each dataset. Finally, the Wilcoxon rank sum test was employed to investigate disparities between the two domains and assess whether significant differences exist between them.

### *3.1. Data Description*

Datasets are extracted from two publicly available repositories that are widely used, containing hundreds of classified datasets: (i) UCI Machine Learning Repository, and (ii) Journal Data in Brief. Appendix 1 provides information about used datasets, along with source references. Datasets from 1 to 33 are categorized as business datasets, whereas datasets from 34 to 46 are categorized as education datasets.

### *3.2. MF Description*

MF were computed using the Python Meta-Feature Extractor (PyMFE) package.

This package offers a comprehensive collection of MF proposed in recent literature, facilitating their extraction. Further details about meta-feature extractor (MFE) packages, available in both Python and R, can be explored in [1].

This package provides the following MF groups:

a. General: Encompasses basic measures that offer general aspects of the datasets, including metrics like “the number of attributes and instances" [46].
b. Statistical: Involves measures that capture the “statistical properties of the data, providing insights into data distribution: average, standard deviation, correlation, and kurtosis” [46].
c. Information-theoretic: “Incorporates measures from the information-theory field, based on entropy. These measures assess the amount of information in the data and its complexity” [46].
d. Model-based: Encompasses “measures designed to extract characteristics from predictive learning models. Often based on decision tree (DT) model properties” [46], they are referred to as DT-based MF and may also be induced by other MLA models.
e. Landmarking: Involves “measures based on the performance of a set of fast and simple learning algorithms” [46]. These measures characterize supervised problems and are indirectly derived from the dataset.
f. Clustering: Encompasses measures related to the extraction of information about the dataset using internal and external validation indices. “Internal indices only consider computed clusters, while external indices require class values to assess partition quality” [46].
g. Concept: Focuses on estimating “the variability of class labels among examples and their density” [46].
h. Itemset: Involves “characterizing binary item sets that capture the distribution of values for both single attributes (*one_itemset*) and pairs of attributes (*two_itemset*)” [46].
i. Complexity: Aims “to estimate the difficulty in separating data points into their expected classes” [46]. The complete survey of the complexity measures can be found in [39].

In the M-L literature, the initial three groups outlined earlier are considered the most prevalent and conventional approaches for data characterization. The fourth and fifth groups rely on MLA to derive model complexity or performance measures. The remaining groups are not widely employed in M-L, primarily due to high computational complexity or domain bias. Nevertheless, they may prove valuable in specific learning scenarios or M-L problems [46].

In this analysis, the MF shown in the first column of Table 1 were calculated.

Their definitions are given in previous papers [2], [4], [5], [6], [8], [13], [14], [19], [21], [22]. [23], [24], [26], [30], [33], [34], [35], [37], [39], [40], [42] [43], [44], [45], [46], [48], [49], [50], [53], [55], [56]. Mean and respective standard deviations were used for aggregation form for the MF.

### *3.3. Methods*

The Wilcoxon rank sum test, a nonparametric statistical test method in the field of statistics, was used in this study. The details on the test can be found in [51]. The Wilcoxon rank sum test was applied because:

(i) samples in the two collections do not follow the normal distribution,

(ii) samples from the two collections are of varying lengths.

The test was used to investigate if there were significance based on the p-value. The value of p = 0.05 is set up as a boundary of significance in this research.

Two approximations are usually applied in Wilcoxon test statistics, normal and the chi-square, both of them using significance at a p-value of 0.05. The conclusion drawn was that there exists a noteworthy difference in the meta-feature values between domains, and a disparity in the meta-feature means is observed depending on the domain when such significance is attained. The normal and chi-square tests are based on the asymptotic distributions of the test statistics.

## 4. Research Results and Discussion

Research analysis included 46 datasets: 33 categorized as business domain datasets, and 13 as educational domain datasets. For each of the datasets, MF were calculated, leading to a total of 90 MF. From the initial set of MF, 3 general MF were excluded: *nr_cat*, *cat_to_num*, *num_to_cat*, since there are no categorical features in the datasets. 6 general MF, 39 statistical MF, 24 model MF, 4 information-theory MF, 8 cluster MF, 4 concept MF, 4 itemset MF, and 3 complexity MF were used in this study.

To evaluate if differences were found, a Wilcoxon test was performed. Table 1 provides the results of the Wilcoxon test statistics.

Test results revealed statistically significant differences in 61 MF among data from two domains. Those MF describe educational and business datasets in the same manner and can be used for both domains to develop a proficient meta-model capable of recommending the most appropriate MLA.

*Table 1. Page layout description*

| | Domain | | Wilcoxon test | |
|---|---|---|---|---|
| **Meta-feature** | EM | BM | *Z* | *p* |
| ***General MF*** | | | | |
| *nr_attr* | 26,54 | 22,30 | 0,95 | 0,3411 |
| *nr_bin* | 32,81 | 19,83 | 3,00 | 0.0027** |
| *nr_inst* | 20,52 | 31,08 | 2,39 | 0.0168* |
| *nr_num* | 26,54 | 22,30 | 0,95 | 0,3411 |
| *attr_to_inst* | 25,08 | 22,88 | 0,49 | 0,6256 |
| *inst_to_attr* | 30,62 | 20,70 | 2,44 | 0.0248* |
| ***Statistical MF*** | | | | |
| *Cor.mean* | 30,08 | 18,50 | 2,76 | 0.0057** |
| *Cor.sd* | 28,38 | 19,23 | 2,18 | 0.0291* |
| *Cov.mean* | 35,77 | 18,67 | 3,88 | 0.0001** |
| *Cov.sd* | 35,38 | 18,82 | 3,76 | 0.0002** |
| *eigenvalues.mean* | 33,85 | 19,42 | 3,27 | 0.0011** |
| *eigenvalues.sd* | 33,77 | 19,45 | 3,24 | 0.0012** |
| *g_mean.mean* | 25,31 | 16,48 | 2,31 | 0.0210* |
| *g_mean.sd* | 24,15 | 17,08 | 1,85 | 0,0648 |
| *h_mean.mean* | 25,69 | 17,15 | 2,19 | 0.0258* |
| *h_mean.sd* | 24,88 | 16,70 | 2,14 | 0.0325* |
| *t_mean.mean* | 33,00 | 19,76 | 3,00 | 0.0026** |
| *t_mean.sd* | 32,46 | 19,97 | 2,83 | 0.0047** |
| *iq_range.mean* | 33,62 | 19,52 | 3,20 | 0.0014** |
| *iq_range.sd* | 33,08 | 19,73 | 3,03 | 0.0025** |
| *Kurtosis.mean* | 32,08 | 20,12 | 2,71 | 0.0068** |
| *Kurtosis.sd* | 31,69 | 20,27 | 2,59 | 0.0096** |
| *Mad.mean* | 33,15 | 19,70 | 3,05 | 0.0023** |
| *Mad.sd* | 33,08 | 19,73 | 3,03 | 0.0025** |
| *Max.mean* | 34,31 | 19,24 | 3,42 | 0.0006** |
| *Mean.sd* | 33,00 | 19,76 | 3,00 | 0.0027** |
| *Median.mean* | 32,62 | 19,91 | 2,88 | 0.0040** |
| *Median.sd* | 29,23 | 21,24 | 1,81 | 0,071 |
| *Min.mean* | 29,77 | 21,03 | 1,98 | 0.0479* |
| *Min.sd* | 27,00 | 22,12 | 1,10 | 0,2698 |
| *nr_cor_attr* | 29,15 | 21,27 | 1,78 | 0,0729 |
| *nr_norm* | 31,54 | 19,55 | 2,57 | 0.0129* |
| *nr_outliers* | 25,77 | 22,61 | 0,71 | 0,4791 |
| *Range.mean* | 34,54 | 19,15 | 3,49 | 0.0005** |
| *Range.sd* | 33,04 | 19,74 | 3,01 | 0.0026** |
| *sd.mean* | 34,38 | 19,21 | 3,44 | 0.0006** |
| *sd.sd* | 33,54 | 19,55 | 3,17 | 0.0015** |
| *var.mean* | 33,85 | 19,42 | 3,27 | 0.00111** |
| *var.sd* | 33,85 | 19,42 | 3,27 | 0.00111** |
| *skewness.mean* | 20,46 | 24,70 | -0,95 | 0,3414 |
| *skewness.sd* | 13,62 | 27,39 | -3,12 | 0.0018** |
| *Sparsity.mean* | 26,58 | 22,29 | 0,97 | 0,3352 |
| *Sparsity.sd* | 26,54 | 22,30 | 0,95 | 0,3413 |
| ***Information - theory*** | | | | |
| *attr_conc.mean* | 34,54 | 19,15 | 3,49 | 0.0005** |
| *attr_conc.sd* | 31,23 | 20,45 | 2,44 | 0.0147* |
| *attr_ent.mean* | 22,08 | 24,06 | -0,44 | 0,6605 |
| *attr_ent.sd* | 26,70 | 22,23 | 1,01 | 0,3113 |
| ***Model based*** | | | | |
| *leaves* | 36,50 | 18,38 | 4,12 | 0.0001** |
| *leaves_branch.mean* | 35,96 | 18,59 | 3,95 | 0.0001** |
| *leaves_branch.sd* | 35,96 | 18,59 | 3,95 | 0.0001** |
| *leaves_corrob.mean* | 36,62 | 18,33 | 4,15 | 0.0001** |
| *leaves_corrob.sd* | 23,23 | 23,61 | -0,07 | 0,9417 |
| *leaves_homo.mean* | 33,81 | 19,44 | 3,26 | 0.0011** |
| *leaves_homo.sd* | 23,19 | 23,62 | -0,09 | 0,9318 |
| *leaves_per_class.mean* | 20,42 | 24,71 | -0,98 | 0,3277 |

| | | | | |
|---|---|---|---|---|
| *leaves_per_class.sd* | 21,08 | 24,45 | -0,76 | 0,4489 |
| *nodes* | 36,50 | 18,38 | 4,12 | 0.0001** |
| *nodes_per_attr* | 36,70 | 18,28 | 4,19 | 0.0001** |
| *nodes_per_inst* | 35,92 | 18,61 | 3,93 | 0.0001** |
| *nodes_per_level.mean* | 36,81 | 18,26 | 4,22 | 0.0001** |
| *nodes_per_level.sd* | 35,54 | 18,44 | 3,86 | 0.0001** |
| *nodes_repeated.mean* | 36,65 | 18,32 | 4,16 | 0.0001** |
| *nodes_repeated.sd* | 31,14 | 16,47 | 3,53 | 0.0004** |
| *tree_depth.mean* | 35,96 | 18,59 | 3,95 | 0.0001** |
| *tree_depth.sd* | 36,12 | 18,53 | 4,00 | 0.0001** |
| *tree_imbalance.mean* | 19,96 | 24,89 | -1,11 | 0,2662 |
| *tree_imbalance.sd* | 14,46 | 26,11 | -2,62 | 0.0087** |
| *tree_shape.mean* | 17,27 | 25,95 | -1,97 | 0.0491* |
| *tree_shape.sd* | 34,88 | 19,02 | 3,60 | 0.0003** |
| *var_importance.mean* | 22,92 | 23,73 | -0,17 | 0,8643 |
| *var_importance.sd* | 22,23 | 24,00 | -0,39 | 0,6962 |
| ***Cluster*** | | | | |
| *ch* | 35,77 | 18,67 | 3,88 | 0.0001** |
| *int* | 35,86 | 18,64 | 3,90 | 0.0001** |
| *nre* | 36,62 | 18,33 | 4,15 | 0.0001** |
| *pb* | 16,54 | 23,72 | -1,74 | 0,0816 |
| *sc* | 26,50 | 22,32 | 1,03 | 0,3049 |
| *sil* | 32,08 | 19,31 | 2,94 | 0.0033** |
| *vdb* | 35,77 | 18,67 | 3,88 | 0.0001** |
| *vdu* | 25,75 | 17,44 | 2,09 | 0.0368* |
| ***Concept*** | | | | |
| *wg_dist.mean* | 18,65 | 19,19 | -0,13 | 0,8979 |
| *wg_dist.sd* | 19,77 | 18,58 | 0,30 | 0,7624 |
| *Cohesiveness.mean* | 23,50 | 16,56 | 1,85 | 0,065 |
| *Cohesiveness.sd* | 24,96 | 15,77 | 2,45 | 0.0143* |
| ***Complexity*** | | | | |
| *t2* | 25,08 | 22,88 | 0,49 | 0,6256 |
| *t3* | 35,92 | 18,61 | 3,93 | 0.0001** |
| *t4* | 28,46 | 21,55 | 1,56 | 0,1183 |
| ***Itemset*** | | | | |
| *one_itemset.mean* | 30,62 | 20,70 | 2,24 | 0.0248* |
| *one_itemset.sd* | 19,27 | 25,17 | -1,33 | 0,1836 |
| *two_itemset.mean* | 26,54 | 22,30 | 0,95 | 0,3413 |
| *two_itemset.sd* | 4,86 | 26,91 | -2,73 | 0.0063** |

** significant at p<.05, ** significant at p<.01*

Hereinafter, MF that have universal values at the social science domain level will be described, attempting to cover the main properties of social science data.

Research results revealed similar patterns in the educational and business data concerning the following MF.

a. general (*nr_attr*, *nr_num*, *attr_to_inst*) - The meta-feature number of attributes and derived MF number of numerical attributes and ratio of attributes to instances characterize the complexity of the given task. From the perspective of complexity, education and business datasets are similar, and those 3 MFs directly address the curse of dimensionality issue.
b. Statistical (*g_mean.sd*, *Median.sd*, *Min.sd*, *nr_cor_attr*, *nr_outliers*, *skewness.mean*, *Sparsity.mean*, *Sparsity.sd*) - The geometric mean and median are mean values that are less affected by outliers. Those measures are equal in cases where there is an exact consistent multiplicative relationship between all numbers). It should be taken into account that *g_mean* and median do not differ in standard deviation, while differing in mean value. Datasets usually contain anomalies, also known as outliers, which should be detected and treated properly. Business and education data do not differ in the number of outliers and minimal values (aggregated by standard deviations), as well as minimum values which are strongly related to outliers. The overall correlation between attributes is also the same in education and business data. Skewness refers to a lack of symmetry in probability distribution, determining feature normality that will directly influence the selection of algorithms in terms of parametric or nonparametric choice. Sparsity "indicates the degree of discreteness of the values in each attribute" [50]. The ability to generalize it on the level of social sciences leads to simplification of the M-L process.
c. information-theory (*attr_ent.mean*, *attr_ent.sd*) - Entropy determines one of the most important aspects concerning information that attributes bring about the class, tackling the class imbalance challenge. The entropy values for the education and business datasets show no differences, being to conclude that most attributes carry an equal amount of information. As stated earlier, there are no differences in business and education data regarding skewness. It is important to note that skewness and entropy are related, since a skewed distribution would mean low entropy and vice versa. So, it is possible to conclude that the presented results are consistent.
d. model-based (*leaves_corrob.sd*, *leaves_homo.sd*, *leaves_per_class.mean*, *leaves_per_class.sd*, *tree_imbalance.mean*, *var_importance.mean*, *var_importance.sd*) - Average leaf corroboration quantifies the average strength of support for each tree leaf, with support measured by the number of training instances corresponding to the paths terminating in each leaf. This descriptor aims to gauge the level of support received by each element of the tree from the sample [6]. Leaves homogeneity refers to the ratio of the number of leaves to the tree's shape. It illustrates the distribution of leaves within the tree, reflecting the extent of attribute label correlations for the given task. [6] Leaves-related measures are indicative of model performance. Similar patterns in education and business data for these MF indicate similar concept complexities in the structure of both domain datasets. A balanced tree indicates that no leaf nodes are distanced from the root.

Variable importance meta-feature shows no difference both in terms of mean value and standard deviation. This factor tackles feature informativeness [14], which shows similarities. The variable importance technique considers the correlation structure of the MF [59], being in line with the results of the statistical meta-feature correlation.

e. cluster (*pb*, *sc*) - *Pb* meta-feature computes the correlation between class matching and instance distances [37]. Also, it refers to the correlation that indicates once again similarity in this aspect between education and business data. *Sc* meta-feature computes the number of clusters with a size smaller than a given size. [44]
f. concept (*wg_dist.sd*, *Cohesiveness.mean*) - Concept MF were found to be related. Cohesiveness is similar to the weighted average (*wg_dist*) used for class variation, nevertheless, attends exclusively to the number of examples. [56]
g. complexity (*t2*, *t4*) - Regarding *t2*, it reflects the data sparsity, while *t4* "gives a rough measure of the proportion of relevant dimensions for the dataset" [39]. Both dimensions sparsity and correlation were previously found to be similar, proving the consistency of the results.
h. itemset (*one_itemset.sd*, *two_itemset.mean*) - The pattern information provided by a one-item set explicitly conveys the information of each attribute individually. Conversely, the two-item set offers correlation information regarding pairs of attributes. Together, they describe complementary aspects of the dataset [53].

When considering their impact on the behavior of MLA, these MF can be handled equivalently in both educational and business data analysis. However, there are statistically significant differences between education and business data in other 61 MF, which should be taken into account when developing meta-models in these two domains. The diversity of social science data among two domains, when looking at most of the measured MF, means that nowadays it is imperative to examine domain specificity of data characteristics. This is especially important regarding social data, which are becoming more and more dynamic. The speed at which businesses and educational institutions move these days, with ever-faster engagements and transactions requires an in-depth analysis of domain data.

## 5. Conclusion

M-L is determined by numerous aspects, such as data properties measured by MF, or hyperparameters optimization, among others. To achieve the success of M-L, the speed and explainability of the different aspects are crucial. This paper contributes to this matter by studying social science data MF. The main research problem in this paper was to extract meta-feature values for social science data, identifying differences in meta-feature values among business and educational datasets.

The response to the first research question is as follows: "The MF of education data exhibit high values across the majority of the measured MF." This statement also addresses the one segment of the second research question. The second question is answered with the following conclusion: "Business data have lower values for most of the MF".

Regarding the third question, the conducted experiments suggest that the most crucial features vary depending on the domain of origin. Thus, it can be concluded: "There are differences in all eight groups of MF between educational and business data."

To the best of the authors' knowledge, this is the first paper that addresses the domain specificity of social science data and examines their characteristics in terms of MF. A thorough study of what are MF of datasets from the educational and business arena was provided, with the extraction of 8 sets of MF for education and business datasets. This research is needed to identify common aspects of both domains and may be significant in defining the topology of the dataset space. The solution to this problem is to use common MF on the social data level, while extracting subdomain-specific data properties for the different MF.

Scientific contributions of this research are:

a. systematic exploration of the huge number of MF that can explain social science data. These features will become the predictive features in the meta-models;
b. the increased explainability of education and business data, as well as the improved speed of the M-L process through characteristics measured by MF, can lead to restricting the search in given configuration space for M-L;
c. an empirical comparison of education and business data properties.

The research findings contribute to a more profound understanding of social data, and the identified differences between datasets are expected to enhance the application of MLA in this context.

Naturally, the results are constrained by the utilized data and MF. Future work aims to replicate the presented analysis scheme on a larger scale of data within the social sciences domain. The approach will be expanded to tackle multi-class problems and nominal attributes, as there remain numerous open issues to address. Employing a broader array of datasets may lead to increased generalization, thereby enhancing the interpretability of the results.

**Acknowledgements**

*This work has been fully supported by Croatian Science Foundation under the project UIP-2020-02-6312.*

**References:**

[1]. Alcobaça, E., Siqueira, F., Rivolli, A., Garcia, L. P., Oliva, J. T., & De Carvalho, A. C. (2020). MFE: Towards reproducible meta-feature extraction. *The Journal of Machine Learning Research*, *21*(1), 4503-4507.

[2]. Alexandros, K., & Melanie, H. (2001). Model selection via meta-learning: a comparative study. *International Journal on Artificial Intelligence Tools*, *10*(4), 525-554.

[3]. Kalousis, A., & Theoharis, T. (1999). Noemon: Design, implementation and performance results of an intelligent assistant for classifier selection. *Intelligent Data Analysis*, *3*(5), 319-337.

[4]. Ali, S., & Smith-Miles, K. A. (2006). A meta-learning approach to automatic kernel selection for support vector machines. *Neurocomputing*, *70*(1-3), 173-186. Doi: 10.1016/j.neucom.2006.03.004

[5]. Ali, S., & Smith, K. A. (2006). On learning algorithm selection for classification. *Applied Soft Computing Journal, 6*(2), 119–138.
Doi: 10.1016/j.asoc.2004.12.002

[6]. Bensusan, H., Giraud-Carrier, C. G., & Kennedy, C. J. (2000). A Higher-order Approach to Meta-learning. *ILP Work-in-progress reports*, *35*, 44.

[7]. Bergmann, H., Mueller, S., & Schrettle, T. (2014). The use of global entrepreneurship monitor data in academic research: A critical inventory and future potentials. *International Journal of Entrepreneurial Venturing*, *6*(3), 242-276.

[8]. Bezdek, J. C., & Pal, N. R. (1998). Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 28*(3), 301–315. Doi: 10.1109/3477.678624

[9]. Bilalli, B., Abelló Gamazo, A., & Aluja Banet, T. (2017). On the predictive power of meta-features in OpenML. *International Journal of Applied Mathematics and Computer Science*, *27*(4), 697-712.

[10]. Bogatinovski, J., Todorovski, L., Džeroski, S., & Kocev, D. (2022). Explaining the performance of multilabel classification methods with data set properties. *International Journal of Intelligent Systems*, *37*(9), 6080-6122.

[11]. Brazdil, P. (2008). Christophe giraud carrier, carlos soares, and ricardo vilalta. *Metalearning: Applications to Data Mining*. Springer Science & Business Media.

[12]. Brazdil, P. B., Soares, C., & Da Costa, J. P. (2003). Ranking learning algorithms: Using IBL and meta-learning on accuracy and time results. *Machine Learning*, *50*, 251-277.

[13]. Calinski, T., & Harabasz, J. (1974). Communications in Statistics A dendrite method for cluster analysis. *Communications in Statistics, 3*(1), 1–27.

[14]. Castiello, C., Castellano, G., Fanelli, A.M. (2005). Meta-data: Characterization of Input Features for Meta-learning. In: Torra, V., Narukawa, Y., Miyamoto, S. (eds) *Modeling Decisions for Artificial Intelligence. MDAI 2005. Lecture Notes in Computer Science(), 3558.* Springer, Berlin, Heidelberg.
Doi: 10.1007/11526018_45

[15]. Chen, C., & Shyu, M. L. (2011, August). Clustering-based binary-class classification for imbalanced data sets. In *2011 IEEE International Conference on Information Reuse & Integration*, 384-389. IEEE.

[16]. Chu, X., Wang, J., Li, S., Chai, Y., & Guo, Y. (2022). Empirical study on meta-feature characterization for multi-objective optimization problems. *Neural Computing and Applications*, *34*(19), 16255-16273.

[17]. Cohen-Shapira, N., Rokach, L., Shapira, B., Katz, G., & Vainshtein, R. (2019, November). Autogrd: Model recommendation through graphical dataset representation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 821-830.

[18]. Cui, C., Hu, M., Weir, J. D., & Wu, T. (2016). A recommendation system for meta-modeling: A meta-learning based approach. *Expert Systems with Applications*, *46*, 33-44.

[19]. Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2), 224–227.

[20]. Dessì, N., & Pes, B. (2015). Similarity of feature selection methods: An empirical study across data intensive classification tasks. *Expert Systems with Applications*, *42*(10), 4632-4642.

[21]. Dunn, J. C. (1974). Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics, 4*(1), 95–104. Doi: 10.1080/01969727408546059

[22]. Engels, R., & Theusinger, C. (1998a). Using a data metric for preprocessing advice for data mining applications. In *ECAI*, *98*, 430-434.

[23]. Engels, R., & Theusinger, C. (1998b). Using a Data Metric for Preprocessing Advice for Data Mining Applications. In *ECAI*, *98*, 430-434.

[24]. Feurer, M., Springenberg, J. T., & Hutter, F. (2014). Using meta-learning to initialize Bayesian optimization of hyperparameters. In *Proceedings of the 2014 International Conference on Meta-learning and Algorithm Selection, 1201* (*MLAS'14*).

[25]. Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. *Automated machine learning: Methods, systems, challenges*, 3-33.

[26]. Filchenkov, A., & Pendryak, A. (2015). Datasets meta-feature description for recommending feature selection algorithm. *2015 Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT)*, 11–18.

[27]. Garouani, Moncef et al. (2022) Using Meta-Learning for Automated Algorithms Selection and Configuration: An Experimental Framework for Industrial Big Data. *Journal of Big Data 9*(1).
Doi: 10.1186/s40537-022-00612-4

[28]. Gómez Guillén, D., & Rojas Espinosa, A. (2017). A meta-analysis on classification model performance in real-world datasets: an exploratory view. *Applied Artificial Intelligence*, *31*, 715-732.

[29]. Kanda, J., De Carvalho, A., Hruschka, E., Soares, C., & Brazdil, P. (2016). Meta-learning to select the best meta-heuristic for the traveling salesman problem: A comparison of meta-features. *Neurocomputing, 205*, 393-406.

[30]. Kalousis, A., & Theoharis, T. (1999). NOEMON: Design, implementation and performance results of an intelligent assistant for classifier selection. *Intelligent Data Analysis, 3*(5), 319–337. Doi: 10.3233/IDA-1999-3502

[31]. Kiang, M. Y. (2003). A comparative assessment of classification methods. *Decision support systems*, *35*(4), 441-454.

[32]. King, R. D., Feng, C., & Sutherland, A. (1995). Statlog: comparison of classification algorithms on large real-world problems. *Applied Artificial Intelligence an International Journal*, *9*(3), 289-333.

[33]. Köpf, C., & Iglezakis, I. (2002). Combination of task description strategies and case base properties for meta-learning. In *Proceedings of the 2nd international workshop on integration and collaboration aspects of data mining, decision support and meta-learning*, 65-76.

[34]. Köpf, C., Taylor, C., & Keller, J. (2000, January). Meta-analysis: From data characterisation for meta-learning to meta-regression. In *Proceedings of the PKDD-00 workshop on data mining, decision support, meta-learning and ILP*.

[35]. Kuba, P., Brazdil, P., Soares, C., & Woznica, A. (2002). Exploiting sampling and meta-learning for parameter setting for support vector machines. *8th IBERAMIA Workshop on Learning and Data Mining*.

[36]. Kwon, O., & Sim, J. M. (2013). Effects of data set features on the performances of classification algorithms. *Expert Systems with Applications*, *40*(5), 1847-1857.

[37]. Lev, J. (1949). The Point Biserial Coefficient of Correlation. *The Annals of Mathematical Statistics, 20*(1), 125–126. Doi: 10.1214/aoms/1177730103

[38]. Li, X., Wang, T., & Wang, H. (2017). Exploring n-gram features in clickstream data for MOOC learning achievement prediction. In *Database Systems for Advanced Applications: DASFAA 2017 International Workshops: BDMS, BDQM, SeCoP, and DMMOOC, Suzhou, China, March 27-30, 2017, Proceedings 22*, 328-339. Springer International Publishing.

[39]. Lorena, A. C., Garcia, L. P., Lehmann, J., Souto, M. C., & Ho, T. K. (2019). How complex is your classification problem? a survey on measuring classification complexity. *ACM Computing Surveys (CSUR)*, *52*(5), 1-34.

[40]. Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (1994). Machine Learning, Neural and Statistical Classification. *Technometrics, 37*(4), 459. Doi: 10.2307/1269742

[41]. Monteiro, J. P., Ramos, D., Carneiro, D., Duarte, F., Fernandes, J. M., & Novais, P. (2021). Meta-learning and the new challenges of machine learning. *International Journal of Intelligent Systems*, *36*(11), 6240-6272.

[42]. Nock, R., & Jappy, P. (1999). Decision tree based induction of decision lists. *Intelligent Data Analysis*, *3*(3), 227-240. Doi: 10.1016/s1088-467x(99)00020-7

[43]. Peng, Y., Wang, G., Kou, G., & Shi, Y. (2011). An empirical study of classification algorithm evaluation for financial risk prediction. *Applied Soft Computing*, *11*(2), 2906-2915.

[44]. Pimentel, B. A., & De Carvalho, A. C. (2019). A new data characterization for selecting clustering algorithms using meta-learning. *Information Sciences*, *477*, 203-219. Doi: 10.1016/j.ins.2018.10.043

[45]. Reif, M., Shafait, F., Goldstein, M., Breuel, T., & Dengel, A. (2014). Automatic classifier selection for non-experts. *Pattern Analysis and Applications, 17*(1), 83–96. Doi: 10.1007/s10044-012-0280-z

[46]. Rivolli, A., Garcia, L. P. F., Soares, C., Vanschoren, J., & de Carvalho, A. C. P. L. F. (2022). Meta-features for meta-learning. *Knowledge-Based Systems, 240*, 108101. Doi: 10.1016/j.knosys.2021.108101

[47]. Romero, C., Olmo, J. L., & Ventura, S. (2013). A meta-learning approach for recommending a subset of white-box classification algorithms for Moodle datasets. In *Educational Data Mining 2013*.

[48]. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65. Doi: 10.1016/0377-0427(87)90125-7

[49]. Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(1), 73–79. Doi: 10.1002/widm.2

[50]. Salama, M. A., Hassanien, A. E., & Revett, K. (2013). Employment of neural network and rough set in meta-learning. *Memetic Computing, 5*(3), 165–177. Doi: 10.1007/s12293-013-0114-6

[51]. Scheff, S. W. (2016). *Fundamental statistical principles for the neurobiologist: A survival guide*. Academic Press.

[52]. Sivakumar, S., Nayak, S. R., Vidyanandini, S., Kumar, J. A., & Palai, G. (2018). An empirical study of supervised learning methods for breast cancer diseases. *Optik, 175*, 105–114. Doi: 10.1016/j.ijleo.2018.08.112

[53]. Song, Q., Wang, G., & Wang, C. (2012). Automatic recommendation of classification algorithms based on data set characteristics. *Pattern Recognition, 45*(7), 2672–2689. Doi: 10.1016/j.patcog.2011.12.025

[54]. Cunha, T., Soares, C., & de Carvalho, A. C. (2018). Metalearning and Recommender Systems: A literature review and empirical study on the algorithm selection problem for Collaborative Filtering. *Information Sciences*, *423*, 128-144.

[55]. Vilalta, R. (1999). Understanding Accuracy Performance Through Concept Characterization and Algorithm Analysis. *Workshop on Recent Advances in Meta-Learning and Future Work, 16th International Conference on Machine Learning*, 3–9.

[56]. Vilalta, R., & Drissi, Y. (2002). A characterization of difficult problems in classification. *Proceedings of the International Conference on Machine Learning and Applications*.

[57]. Vilalta, R., Giraud-Carrier, C.G., Brazdil, P., Soares, C.: Using meta-learning to support data mining. *Int. J. Comput. Sci. Appl. 1*(1), 31–45 (2004)

[58]. Wang G, Song Q, Zhang X, Zhang K (2014) a generic multilabel learning-based classification algorithm recommendation method. *Acm Trans Knowl Discov Data (TKDD) 9*(1), 1–30

[59]. Woźnica, K., & Biecek, P. (2021). Towards explainable meta-learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 505-520. Springer, Cham.

[60]. Wu, M. S., & Lu, J. Y. (2018). Automated Machine Learning Algorithm Mining for Classification Problem. In *Machine Learning and Data Mining in Pattern Recognition: 14th International Conference, MLDM 2018, New York, NY, USA, July 15-19, 2018, Proceedings, Part I 14*, 380-392. Springer International Publishing.

[61]. Zhang, X., Li, R., Zhang, B., Yang, Y., Guo, J., & Ji, X. (2019). An instance-based learning recommendation algorithm of imbalance handling methods. *Applied Mathematics and Computation, 351*, 204-218.

**Appendix 1**

| | **Title of the dataset** | **Link** | **References** |
|---|---|---|---|
| **1** | Company Bankruptcy Prediction | UCI Machine Learning Repository: Taiwanese Bankruptcy Prediction Data Set | Liang D., & Tsai C.F. (2020). *UCI Machine Learning Repository*. Retrieved from http://archive.ics.uci.edu/ml |
| **2** | Wholesale customers Data Set | UCI Machine Learning Repository: Wholesale customers Data Set | Cardoso, Margarida G. M. S. (2014). *UCI Machine Learning Repository*. Retrieved from http://archive.ics.uci.edu/ml |
| **3** | Data for the ins and outs of involuntary part-time employment all_baseline.xls | https://www.sciencedirect.com/science/article/pii/S2352340920315651 | Borowczyk-Martins, D., & Lalé, E. (2021). Data for the ins and outs of involuntary part-time employment. *Data in Brief*, *34*, 106686. https://doi.org/10.1016/J.DIB.2020.106686 |
| **4** | Data for the ins and outs of involuntary part-time employment all_reclassified.xls | https://www.sciencedirect.com/science/article/pii/S2352340920315651 | Borowczyk-Martins, D., & Lalé, E. (2021). Data for the ins and outs of involuntary part-time employment. *Data in Brief*, *34*, 106686. https://doi.org/10.1016/J.DIB.2020.106686 |
| **5** | Dataset on the perceptions of ordinary people on the persistence of bribery practices in Nigeria | https://www.sciencedirect.com/science/article/pii/S2352340920314967 | Sani, A. S., & Abu Bakar, A. S. B. (2021). Dataset on the perceptions of ordinary people on the persistence of bribery practices in Nigeria. *Data in Brief*, *34*, 106616. https://doi.org/10.1016/J.DIB.2020.106616 |
| **6** | Survey data of coronavirus (COVID-19) thought concern, employees' work performance, employees background, feeling about job, work motivation, job satisfaction, psychological state of mind and family commitment in two middle east countries | https://www.sciencedirect.com/science/article/pii/S235234092031533X | Mgammal, M. H., & Al-Matari, E. M. (2021). Survey data of coronavirus (COVID-19) thought concern, employees' work performance, employees background, feeling about job, work motivation, job satisfaction, psychological state of mind and family commitment in two middle east countries. *Data in Brief*, *34*, 106661. https://doi.org/10.1016/J.DIB.2020.106661 |
| **7** | Data from an incentivized laboratory experiment on strategic medical choices | https://www.sciencedirect.com/science/article/pii/S2352340921002109 | Ge, G., & Godager, G. (2021). Data from an incentivized laboratory experiment on strategic medical choices. *Data in Brief*, *35*, 106926. https://doi.org/10.1016/J.DIB.2021.106926 |
| **8** | Hedonic dataset of the metropolitan housing market – Cases in South Korea prices_busan.xlsx | https://www.sciencedirect.com/science/article/pii/S235234092100161X | Song, Y., Ahn, K., An, S., & Jang, H. (2021). Hedonic dataset of the metropolitan housing market – Cases in South Korea. *Data in Brief*, *35*, 106877. https://doi.org/10.1016/J.DIB.2021.106877 |
| **9** | Hedonic dataset of the metropolitan housing market – Cases in South Korea prices_daegu.xlsx | https://www.sciencedirect.com/science/article/pii/S235234092100161X | Song, Y., Ahn, K., An, S., & Jang, H. (2021). Hedonic dataset of the metropolitan housing market – Cases in South Korea. *Data in Brief*, *35*, 106877. https://doi.org/10.1016/J.DIB.2021.106877 |
| **10** | Hedonic dataset of the metropolitan housing market – Cases in South Korea prices_daejeon.xlsx | https://www.sciencedirect.com/science/article/pii/S235234092100161X | Song, Y., Ahn, K., An, S., & Jang, H. (2021). Hedonic dataset of the metropolitan housing market – Cases in South Korea. *Data in Brief*, *35*, 106877. https://doi.org/10.1016/J.DIB.2021.106877 |
| **11** | Hedonic dataset of the metropolitan housing market – Cases in South Korea prices_gwangju.xlsx | https://www.sciencedirect.com/science/article/pii/S235234092100161X | Song, Y., Ahn, K., An, S., & Jang, H. (2021). Hedonic dataset of the metropolitan housing market – Cases in South Korea. *Data in Brief*, *35*, 106877. https://doi.org/10.1016/J.DIB.2021.106877 |

| | | | |
|---|---|---|---|
| **12** | Retail customers' satisfaction with banks in Greece: A multicriteria analysis of a dataset | https://www.sciencedirect.com/science/article/pii/S2352340921001992 | Drosos, D., Skordoulis, M., Tsotsolas, N., Kyriakopoulos, G. L., Gkika, E. C., & Komisopoulos, F. (2021). Retail customers' satisfaction with banks in Greece: A multicriteria analysis of a dataset. *Data in Brief*, *35*, 106915. https://doi.org/10.1016/J.DIB.2021.106915 |
| **13** | A dataset of factors influencing consumer behavior towards bringing own shopping bags instead of using plastic bags in Vietnam | https://www.sciencedirect.com/science/article/pii/S2352340921005102?via%3Dihub | Nguyen, T. P. L. (2021). A dataset of factors influencing consumer behavior towards bringing own shopping bags instead of using plastic bags in Vietnam. Data in Brief, 37, 107226. https://doi.org/10.1016/J.DIB.2021.107226 |
| **14** | Digital adoption by enterprises in Malaysian industrial sectors during COVID-19 pandemic: A data article | https://www.sciencedirect.com/science/article/pii/S2352340921004819 | Muhamad, S., Kusairi, S., Man, M., Majid, N. F. H., & Wan Kassim, W. Z. (2021). Digital adoption by enterprises in Malaysian industrial sectors during COVID-19 pandemic: A data article. *Data in Brief*, *37*, 107197. https://doi.org/10.1016/J.DIB.2021.107197 |
| **15** | Dataset on social capital and knowledge integration in project management | https://www.sciencedirect.com/science/article/pii/S235234092030127X | Ekemen, M. A., & Şeşen, H. (2020). Dataset on social capital and knowledge integration in project management. *Data in Brief*, *29*, 105233. https://doi.org/10.1016/J.DIB.2020.105233 |
| **16** | Data modelling consumer-generated content usage for apparel shopping | https://www.sciencedirect.com/science/article/pii/S235234092030929X | Tobias-Mamina, R. J., & Kempen, E. (2020). Data modelling consumer-generated content usage for apparel shopping. *Data in Brief*, *31*, 106035. https://doi.org/10.1016/J.DIB.2020.106035 |
| **17** | From selected multi-sensory dimensions to positive word of mouth: Data on what really drives generation z consumers to be attached to quick service restaurants in bloemfontein, South Africa? | https://www.sciencedirect.com/science/article/pii/S2352340920311732 | Maziriri, E. T., Rukuni, T. F., & Chuchu, T. (2020). From selected multi-sensory dimensions to positive word of mouth: Data on what really drives generation z consumers to be attached to quick service restaurants in bloemfontein, south africa? *Data in Brief*, *32*, 106279. https://doi.org/10.1016/J.DIB.2020.106279 |
| **18** | A survey dataset on determinants of administrative corruption | https://www.sciencedirect.com/science/article/pii/S2352340919311758 | Al-Jundi, S. A. (2019). A survey dataset on determinants of administrative corruption. *Data in Brief*, *27*, 104820. https://doi.org/10.1016/J.DIB.2019.104820 |
| **19** | Data to model the influence of CSR on consumer behaviors: A process approach | https://www.sciencedirect.com/science/article/pii/S2352340919310686 | Castro-González, S., Bande, B., Fernández-Ferrín, P., & Kimura, T. (2019). Data to model the influence of CSR on consumer behaviors: A process approach. *Data in Brief*, *27*, 104713. https://doi.org/10.1016/J.DIB.2019.104713 |
| **20** | The brief data of the relation of living as commuters and quality of life | https://www.sciencedirect.com/science/article/pii/S2352340919308959 | Rahmadana, M. F., & Sagala, G. H. (2019). The brief data of the relation of living as commuters and quality of life. *Data in Brief*, *26*, 104540. https://doi.org/10.1016/J.DIB.2019.104540 |
| **21** | The dataset for validation of customer inspiration construct in Malaysian context | https://www.sciencedirect.com/science/article/pii/S2352340919304858 | Ghouri, A. M., Kin, T. M., Yunus, N. K. bin Y., & Akhtar, P. (2019). The dataset for validation of customer inspiration construct in Malaysian context. *Data in Brief*, *25*, 104131. https://doi.org/10.1016/J.DIB.2019.104131 |
| **22** | Data on patient's satisfaction from an emergency department: Developing strategies with the Multicriteria Satisfaction Analysis | https://www.sciencedirect.com/science/article/pii/S2352340918312587 | Manolitzas, P., Kostagiolas, P., Grigoroudis, E., Intas, G., & Stergiannis, P. (2018). Data on patient's satisfaction from an emergency department: Developing strategies with the Multicriteria Satisfaction Analysis. *Data in Brief*, *21*, 956–961. https://doi.org/10.1016/J.DIB.2018.10.041 |
| **23** | Data on perception of faculty members on the influence of faculty support initiatives on the efficacy of job responsibilities | https://www.sciencedirect.com/science/article/pii/S2352340918307169 | Falola, H. O., Adeniji, A. A., Osibanjo, A. O., Oludayo, O. A., & Salau, O. P. (2018). Data on perception of faculty members on the influence of faculty support initiatives on the efficacy of job responsibilities. *Data in Brief*, *19*, 1594–1599. https://doi.org/10.1016/J.DIB.2018.06.065 |
| **24** | Job design and behavioural outcome of employees in agricultural research training, Ibadan, Nigeria | https://www.sciencedirect.com/science/article/pii/S2352340918307248 | Osibanjo, A. O., Abiodun, A. J., Salau, O. P., Adeniji, A. A., Falola, H. O., & Alimi, I. I. (2018). Job design and behavioural outcome of employees in agricultural research training, Ibadan, Nigeria. *Data in Brief*, *19*, 1880–1887. https://doi.org/10.1016/J.DIB.2018.06.073 |
| **25** | Data on empirical estimation of the relationship between agency costs and ownership structure in Italian listed companies (2002–2013) | https://www.sciencedirect.com/science/article/pii/S2352340918304591 | Rossi, F., Barth, J. R., & Cebula, R. J. (2018). Data on empirical estimation of the relationship between agency costs and ownership structure in Italian listed companies (2002–2013). *Data in Brief*, *18*, 2010–2012. https://doi.org/10.1016/J.DIB.2018.04.106 |
| **26** | Dataset on PGA Tour tournament entry | https://www.sciencedirect.com/science/article/pii/S2352340922001639 | Alegre, I., Canela, M. A., & Pastoriza, D. (2022). Dataset on PGA Tour tournament entry. *Data in Brief*, *41*, 107952. https://doi.org/10.1016/J.DIB.2022.107952 |

| | | | |
|---|---|---|---|
| **27** | A dataset of factors affecting sustainable consumption intention in Vietnam | https://www.sciencedirect.com/science/article/pii/S2352340922003377 | Tran, L. H., Nguyen, N. A., Tran, T. D., & Nguyen, T. P. L. (2022). A dataset of factors affecting sustainable consumption intention in Vietnam. *Data in Brief*, *42*, 108127. https://doi.org/10.1016/J.DIB.2022.108127 |
| **28** | Data modelling of subsistence retail consumer purchase behavior in South Africa | https://www.sciencedirect.com/science/article/pii/S2352340922003043 | Zulu, V. M., & Nkuna, A. M. (2022). Data modelling of subsistence retail consumer purchase behavior in South Africa. *Data in Brief*, *42*, 108094. https://doi.org/10.1016/J.DIB.2022.108094 |
| **29** | Dataset for cognition processes, motivations, spatial presence experience, and customer engagement in retail mobile apps | https://www.sciencedirect.com/science/article/pii/S2352340922004024 | Le, A. N. H., Ho, H. X., Nguyen, D. P., & Cheng, J. M. S. (2022). Dataset for cognition processes, motivations, spatial presence experience, and customer engagement in retail mobile apps. *Data in Brief*, *42*, 108198. https://doi.org/10.1016/J.DIB.2022.108198 |
| **30** | Dataset for understanding the effort and performance of external auditors during the COVID-19 crisis: A remote audit analysis | https://www.sciencedirect.com/science/article/pii/S2352340922003298 | Baatwah, S. R., & Al-Ansi, A. A. (2022). Dataset for understanding the effort and performance of external auditors during the COVID-19 crisis: A remote audit analysis. *Data in Brief*, *42*, 108119. https://doi.org/10.1016/J.DIB.2022.108119 |
| **31** | Personality traits and social loafing among employees working in teams at small and medium enterprises: A cultural perspective data from emerging economies | https://www.sciencedirect.com/science/article/pii/S2352340922002967 | Bokhari, S. A. A., & Aftab, M. (2022). Personality traits and social loafing among employees working in teams at small and medium enterprises: A cultural perspective data from emerging economies. *Data in Brief*, *42*, 108085. https://doi.org/10.1016/J.DIB.2022.108085 |
| **32** | Dataset for an analysis of tourism and economic growth: A study of Sri Lanka | https://www.sciencedirect.com/science/article/pii/S2352340916304346 | Kumar, R. R., & Stauvermann, P. J. (2016). Dataset for an analysis of tourism and economic growth: A study of Sri Lanka. *Data in Brief*, *8*, 723–725. https://doi.org/10.1016/J.DIB.2016.06.066 |
| **33** | Residential construction cost: An Italian survey | https://www.sciencedirect.com/science/article/pii/S2352340917300240 | Canesi, R., & Marella, G. (2017). Residential construction cost: An Italian survey. *Data in Brief*, *11*, 231–235. https://doi.org/10.1016/J.DIB.2017.02.005 |
| **34** | Data-set of academic difficulties among students in western Uganda during COVID-19 induced lockdown | https://www.sciencedirect.com/science/article/pii/S2352340921001359 | Meji M, A., Dennison, M. S., & Mustafa, M. M. (2021). Data-set of academic difficulties among students in western Uganda during COVID-19 induced lockdown. *Data in Brief*, *35*, 106851. https://doi.org/10.1016/J.DIB.2021.106851 |
| **35** | School students' perception, attitudes and skills regarding global citizenship-dataset from Vietnam | https://www.sciencedirect.com/science/article/pii/S2352340921004467 | Nguyen, H. L., Dinh, V. H., Hoang, P. H., Luong, V. T., & Le, A. V. (2021). School students' perception, attitudes and skills regarding global citizenship-dataset from Vietnam. *Data in Brief*, *37*, 107162. https://doi.org/10.1016/J.DIB.2021.107162 |
| **36** | Data on factors characterizing the eLearning experience of secondary school teachers and university undergraduate students in Jordan | https://www.sciencedirect.com/science/article/pii/S2352340920312841 | Obidat, A. H., Alquraan, M., & Obeidat, M. H. (2020). Data on factors characterizing the eLearning experience of secondary school teachers and university undergraduate students in Jordan. *Data in Brief*, *33*, 106402. https://doi.org/10.1016/J.DIB.2020.106402 |
| **37** | Digital literacy and e-learning experiences among the pre-service teachers data | https://www.sciencedirect.com/science/article/pii/S235234092030946X | Tomczyk, Ł. (2020). Digital literacy and e-learning experiences among the pre-service teachers data. *Data in Brief*, *32*, 106052. https://doi.org/10.1016/J.DIB.2020.106052 |
| **38** | Datasets linking ethnic perceptions to undergraduate students learning outcomes in a Nigerian Tertiary Institution | https://www.sciencedirect.com/science/article/pii/S2352340918302798 | Badejo, J. A., John, T. M., Omole, D. O., Ucheaga, E. G., Popoola, S. I., Odukoya, J. A., Ajayi, P. O., Aboyade, M., & Atayero, A. A. (2018). Datasets linking ethnic perceptions to undergraduate students learning outcomes in a Nigerian Tertiary Institution. *Data in Brief*, *18*, 760–764. https://doi.org/10.1016/J.DIB.2018.03.069 |
| **39** | The role of gender on academic performance in STEM-related disciplines: Data from a tertiary institution | https://www.sciencedirect.com/science/article/pii/S2352340918302579 | John, T. M., Badejo, J. A., Popoola, S. I., Omole, D. O., Odukoya, J. A., Ajayi, P. O., Aboyade, M., & Atayero, A. A. (2018). The role of gender on academic performance in STEM-related disciplines: Data from a tertiary institution. *Data in Brief*, *18*, 360–374. https://doi.org/10.1016/J.DIB.2018.03.052 |
| **40** | A dataset of the relationship between emotional intelligence and teamwork results of university students | https://www.sciencedirect.com/science/article/pii/S2352340922003535 | Nguyen, T. P. L. (2022). A dataset of the relationship between emotional intelligence and teamwork results of university students. *Data in Brief*, *42*, 108149. https://doi.org/10.1016/J.DIB.2022.108149 |
| **41** | Data survey of students behavioral and psychological adaptations in disaster-prone areas of Mount Merapi in Indonesia | https://www.sciencedirect.com/science/article/pii/S2352340922004231 | Hafida, S. H. N., Isa, N. K. M., Ibrahim, M. H., Jumadi, Toyib, M., & Musiyam, M. (2022). Data survey of students behavioral and psychological adaptations in disaster-prone areas of Mount Merapi in Indonesia. *Data in Brief*, *42*, 108229. https://doi.org/10.1016/J.DIB.2022.108229 |

| | | | |
|---|---|---|---|
| **42** | Dataset exploring organizational culture of K-12 schools | https://www.sciencedirect.com/science/article/pii/S2352340922003833 | Kareem, J., Patrick, H. A., Tantia, V., & Valarmathi B, S. (2022). Dataset exploring organizational culture of K-12 schools. *Data in Brief*, *42*, 108179. https://doi.org/10.1016/J.DIB.2022.108179 |
| **43** | Dataset for integration of sustainability education into the accounting curricula of tertiary education institutions in Jordan | https://www.sciencedirect.com/science/article/pii/S2352340922004279 | Al-Hazaima, H., Al Shbail, M. O., Alshurafat, H., Ananzeh, H., & Al Shbeil, S. O. (2022). Dataset for integration of sustainability education into the accounting curricula of tertiary education institutions in Jordan. *Data in Brief*, *42*, 108224. https://doi.org/10.1016/J.DIB.2022.108224 |
| **44** | Survey dataset for the perceived consciousness towards environmental sustainability by undergraduate students | https://www.sciencedirect.com/science/article/pii/S2352340922001962 | Fuchs, K. (2022). Survey dataset for the perceived consciousness towards environmental sustainability by undergraduate students. *Data in Brief*, *41*, 107985. https://doi.org/10.1016/J.DIB.2022.107985 |
| **45** | Dataset of Factors affecting online cheating by accounting students: The relevance of social factors and the fraud triangle model factors | https://www.sciencedirect.com/science/article/pii/S2352340921010076 | Shbail, M. O. Al, Alshurafat, H., Ananzeh, H., & Al-Msiedeen, J. M. (2022). Dataset of Factors affecting online cheating by accounting students: The relevance of social factors and the fraud triangle model factors. *Data in Brief*, *40*, 107732. https://doi.org/10.1016/J.DIB.2021.107732 |
| **46** | Dataset of international students' acceptance of online distance learning during COVID-19 pandemic: A preliminary investigation | https://www.sciencedirect.com/science/article/pii/S2352340922004346 | Shi-Hui, S., Chaw, L. Y., Aw, E. C. X., & Sham, R. (2022). Dataset of international students' acceptance of online distance learning during COVID-19 pandemic: A preliminary investigation. *Data in Brief*, *42*, 108232. https://doi.org/10.1016/J.DIB.2022.108232 |