www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 17(4), 2026, 1-9. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2026.v17i4.1290

Evaluating the SEMMA Methodology with AutoML Across Phishing Detection, Pollen Classification, and Voice Identification Tasks

Cecilia-Irene Loeza-Mejía[1], *Eddy Sánchez-DelaCruz*[1*]
[1] Artificial Intelligence Lab., Tecnológico Nacional de México/Instituto Tecnológico Superior de Misantla, Veracruz, Mexico.
✉Corresponding author: Sánchez-DelaCruz, E. eddsacx@gmail.com

Abstract. The exponential growth of data has increased the demand for effective machine learning methodologies. The SEMMA (Sample, Explore, Modify, Model, and Assess) methodology stands out for its efficient structure for pattern extraction. Hence, the contribution of this study lies in the application of the SEMMA methodology in phishing detection, pollen classification, and voice identification. Also, the AutoML (automated machine learning) approach was employed to perform model selection, identifying Random Forest for the phishing and voice datasets, and Support Vector Machine for the pollen dataset. Precision values of 0.912, 0.941, and 0.784 were obtained for phishing detection, pollen classification, and voice identification, respectively. The results demonstrate the applicability of SEMMA combined with AutoML on heterogeneous machine learning tasks.
Keywords: SEMMA methodology; automated machine learning; model selection; phishing detection; pollen classification; voice identification; Random Forest; Support Vector Machine.

Article information

Received: Jan 17, 2026
Accepted: May 11, 2026

1 Introduction

Since ancient times, humans have sought to extract knowledge from available information, initially through intuitive reasoning and later through systematic approaches supported by statistics and data analysis methods. In recent decades, technological advances in data acquisition, processing, massive storage, and cloud computing have led to an unprecedented growth in data generated from diverse sources (Fayyad et al., 1996; Zurada & Karwowski, 2011; Li et al., 2019; Holzinger, 2013; Rogalewicz & Sika, 2016). Consequently, the exponential growth of data has driven the need to develop effective methods for extracting useful information.

In this context, machine learning has emerged as a fundamental tool, with applications ranging from search engines to fraud detection systems in financial transactions (Shalev-Shwartz & Ben-David, 2014). Among the various machine learning methodologies, SEMMA stands out as a prominent option. SEMMA consists of five phases: Sample, Explore, Modify, Model, and Assess (SAS Institute Inc., 2003), which provide a clear structure for the knowledge discovery process (Azevedo & Santos, 2008). Despite its origins in the 1990s, SEMMA remains relevant today due to its adaptability to the evolving needs of data analysis and machine learning.

Several studies have validated the applicability of SEMMA in diverse contexts, including educational data mining for student performance prediction (Jadrić et al., 2010; Jacob et al, 2023), sports analytics (Vistro et al., 2019), public health (De-La-Hoz-Correa et al., 2019), resource optimization through IoT systems (López-Torres et al., 2019), and sentiment analysis using unstructured social media data (Hansun et al., 2022). These contributions highlight the versatility of SEMMA for addressing both structured and unstructured data problems across different domains.

In this article, the results previously reported in Loeza-Mejía & Sánchez-DelaCruz (2025) are extended. Unlike the prior study, the experiments were conducted in a different computational environment and expanded through the incorporation of additional experiments on an audio dataset, enabling a broader validation of the methodology under varied technical conditions. Specifically,

three datasets are considered: a dataset for phishing website detection (Vrbančič et al., 2020), a dataset of extracted attributes from pollen images (Tello-Mijares & Flores, 2016), and an audio dataset for multi-voice identification (Duville et al., 2021). These datasets represent a variety of domains and disciplines yet converge in their utility for exploring and validating the applicability of the SEMMA methodology in different data analysis contexts.

The article proceeds with the description of the SEMMA methodology and its application to the case studies in Section 2. Section 3 presents the obtained results, which are subsequently discussed in Section 4. Finally, Section 5 provides the conclusions and outlines possible future research directions.

2 SEMMA Methodology

The SEMMA methodology provides an approach that facilitates both the initial development and continuous maintenance of machine learning projects. This methodology offers a comprehensive structure for the design, development, and evolution of solutions aimed at addressing business problems, while also contributing to the identification and achievement of organizational goals (Azevedo & Santos, 2008; Jadrić et al., 2010). Fig. 1 illustrates the five phases proposed by SAS Institute Inc. (2003), as adapted for the present study.

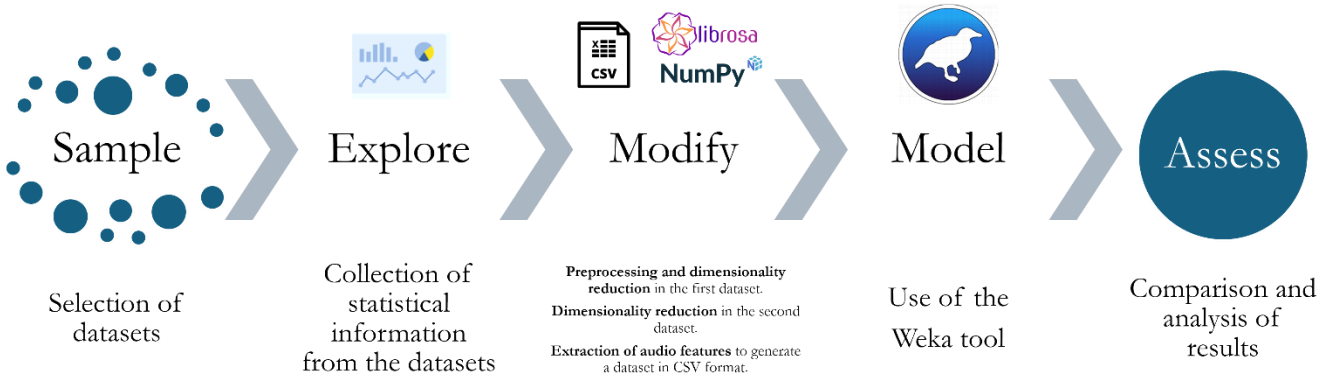

Fig. 1. SEMMA process adapted from (SAS Institute Inc., 2003; Loeza-Mejía et al., 2025)

2.1 Phase 1: Sample

In this phase, the datasets were selected and prepared for analysis. First, the phishing websites detection dataset (Vrbančič et al., 2020) was obtained due to its relevance in the field of cybersecurity (Sarker et al., 2024) and its significance for evaluating the SEMMA methodology in a critical and contemporary context. The dataset[1] provides a series of attributes representing a website, including structural and content-related information from the URL, such as the number of dots, hyphens, slashes, special characters, and the URL length. These numerical attributes are used as features to train and evaluate machine learning algorithms for the binary identification of phishing, i.e., to determine whether a website contains phishing content or not.

Secondly, a dataset containing features extracted from pollen images[2] was obtained. This dataset is relevant to both ecology and public health, where pollen detection and analysis can provide critical insights for the prevention and treatment of allergic and respiratory conditions. The dataset includes a variety of morphological and statistical features such as width, height, perimeter, area, circularity, compactness, as well as multiple texture and intensity statistics. Each row is labeled with its corresponding pollen species, encompassing a total of 12 classes (Tello-Mijares & Flores, 2016). The features include shape descriptors (e.g., width, height, perimeter, and area) and texture measures (e.g., contrast, correlation, energy, and homogeneity), among others.

Lastly, a dataset of Mexican emotional speech (Duville et al., 2021) in WAV format[3] was incorporated. This dataset is particularly relevant to fields such as linguistics, psychology, and sociology, where voice analysis can provide insights into behavioral patterns, cultural differences, and social inequalities.

[1] Available at https://data.mendeley.com/datasets/72ptz43s9v/1
[2] Available at https://doi.org/10.1155/2016/5689346
[3] Available at https://data.mendeley.com/datasets/cy34mh68j9/3

2.2 Phase 2: Explore

During this phase, exploratory data analysis was performed to evaluate the need for preprocessing. The procedure proposed by Zavaleta-Sánchez et al. (2024) was followed, collecting detailed statistical information from datasets. This included record counts for each attribute, as well as calculations of the mean, standard deviation, quartiles, outlier values, and correlation analysis using the Pearson correlation coefficient (Benesty et al., 2009), considering only numerical attributes for its application.

For the first dataset, 1,438 duplicate records were identified, along with 13 attributes exhibiting constant values across all records. In contrast, the second dataset showed no missing values, constant attributes, or duplicate entries. For the audio dataset, provided in WAV format, the process proceeded directly to the *Modify* phase.

2.3 Phase 3: Modify

In this phase, data cleaning and preparation processes were carried out in order to enable subsequent modeling. For the first dataset, 1,438 duplicate records were removed, constant attributes were eliminated, and attributes exhibiting a correlation higher than 90% with other features were excluded. This value was selected as the threshold because the literature has identified correlations above 90% as indicative of a high degree of correlation and potential redundancy between variables (Caruso et al., 2026). After these cleaning and transformation steps, the resulting dataset comprised 87,209 rows and 58 columns, including 57 attribute columns and one class label column. For the second dataset, attributes exhibiting a correlation greater than 90% were also excluded. As a result, the final dataset consisted of 618 rows and 19 columns, including 18 attribute columns and one class label column.

To construct the audio dataset in CSV format, a total of 864 Mexican emotional speech samples in WAV format were collected. The Python libraries *librosa*[4] and *numpy*[5] were used to process the audio data. A variety of audio features were extracted, including chromagram, Mel-frequency cepstral coefficients (MFCCs), zero-crossing rate, spectral bandwidth of order p, dynamic programming beat tracker, tuning, constant-Q chromagram, harmonic and percussive components, spectral contrast, and short-time Fourier transform (STFT), among others. These features were used as attributes, and statistical measures such as mean, mode, median, and interquartile range were calculated. Each extracted feature was converted into a separate column, and a class label column (*child*, *female*, *male*) was added. Additionally, binary encoding was applied to the "corpus" attribute, corresponding to corpus 1, which is composed of 24 nouns and adjectives, and corpus 2, corresponding to 24 words (Duville et al., 2021). Ordinal encoding was also applied to the *word* attribute, corresponding to the word spoken in the audio recording. The resulting dataset comprised 864 rows and 26 columns, including 25 attribute columns and one class label column. It was specifically designed for speaker voice identification (*child*, *female*, or *male*), including acoustic features, audio signal statistics, and processing parameters, along with the corresponding class label.

2.4 Phase 4: Model

In this phase, predictive algorithms for classification (or identification) were used using the data prepared in the previous stages. Weka[6] version 3.8.6 was used to carry out Phases 4 and 5 of the SEMMA methodology. Within Weka, the Auto-Weka (Thornton et al., 2013) method was employed, enabling the automatic selection and optimization of models using default settings. Weka was executed on a Windows 11 Home Single Language 64-bit computer equipped with an Intel(R) Core(TM) i7-1065G7 CPU @ 1.30GHz (8 CPUs) and 12GB RAM.

Algorithms

The following algorithms were automatically selected by AutoML via Auto-Weka based on their superior performance during the validation phase: Random Forest for the phishing and voice datasets, and Support Vector Machine for the pollen dataset.
Random Forest: consists of an ensemble learning method that constructs multiple decision trees using bootstrap samples and random subsets of features. The final prediction is obtained through majority voting among the generated trees, improving classification robustness and reducing overfitting (Breiman, 2001).

[4] https://librosa.org/
[5] https://numpy.org/

Support Vector Machine: consists of a supervised learning algorithm that identifies an optimal separating hyperplane between classes by maximizing the margin between data points of different categories. Kernel functions can be used to model non-linear decision boundaries in complex datasets (Vapnik, 1995; Sánchez-DelaCruz & Loeza-Mejía, 2024).

Algorithm 1 presents the Random Forest procedure, while Algorithm 2 describes the Support Vector Machine process.

Algorithm 1: Random Forest

Input: Training set S, Attribute set X, Class label set Y, Number of trees T
Output: Random forest model RF

Function TrainRandomForest(S, X, Y, T)
    RF ← ∅
    For t = 1 **to** T **do:**
        St ← BootstrapSample(S)
        Xt ← RandomSubset(X)
        DTt ← TrainTree(St, Xt, Y)
        RF ← RF ∪ {DTt}
    End
    Return RF

Function PredictRandomForest(RF, x)
    Predictions ← ∅
    For each DTt ∈ RF **do:**
    yt ← Predict(DTt, x)
    Predictions ← Predictions ∪ {yt}
    End
    Return MajorityVote(Predictions)

Algorithm 2: Support Vector Machine

Input: Training set S, Feature set X, Class label set Y
Output: Trained support vector machine model SVM

Function TrainSVM(S, X, Y)
    Initialize SVM parameters
    Select kernel function K
    Optimize hyperplane parameters using training data S
    Identify support vectors
    Construct SVM model
    Return SVM

Function PredictSVM(SVM, x)
    Compute the decision function using support vectors and kernel K
    Assign class label y to x
    Return y

2.5 Phase 5: Assess

In this phase, the results obtained for each dataset (see Sections 3 and 4) were compared to gain a broader and more generalizable understanding of how the SEMMA methodology can address data analysis problems across various domains and disciplines.

3 Results

Tables 1–3 present the classification results for the phishing, pollen, and voice datasets, respectively. Table 1 shows the results for the phishing dataset, achieving a weighted average TP Rate of 0.908, indicating strong overall classification performance.

Table 2 presents the results for the pollen dataset, which achieved the highest overall performance among the evaluated datasets, with a weighted average TP Rate of 0.940. Several classes, such as Five, Six, Eight, and Ten, obtained near-perfect classification metrics. Finally, Table 3 summarizes the results for the voice dataset. Although the performance was lower compared to the phishing and pollen datasets, the results still demonstrate the feasibility of voice identification using acoustic and statistical audio features, achieving a weighted average TP Rate of 0.785. Among the evaluated classes, the male category obtained the best performance metrics.

Table 1. Results on the phishing dataset

Class	TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area	PRC Area
Positive	0.916	0.096	0.836	0.875	0.804	0.957	0.899
Negative	0.904	0.084	0.953	0.927	0.804	0.957	0.973
Weighted Avg.	**0.908**	**0.088**	**0.912**	**0.909**	**0.804**	**0.957**	**0.947**

Table 2. Results on the pollen dataset

Class	TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area	PRC Area
One	0.957	0.000	1.000	0.978	0.976	0.986	0.971
Two	0.913	0.017	0.808	0.857	0.847	0.980	0.883
Three	0.964	0.007	0.931	0.947	0.942	0.988	0.969
Four	0.838	0.007	0.886	0.861	0.853	0.946	0.875
Five	0.976	0.000	1.000	0.988	0.987	0.999	0.988
Six	0.982	0.000	1.000	0.991	0.990	0.999	0.994
Seven	1.000	0.003	0.941	0.970	0.968	1.000	1.000
Eight	0.981	0.000	1.000	0.990	0.989	0.999	0.991
Nine	0.850	0.017	0.883	0.866	0.847	0.978	0.904
Ten	1.000	0.005	0.955	0.977	0.975	0.999	0.990
Eleven	0.932	0.007	0.944	0.938	0.930	0.997	0.979
Twelve	0.912	0.002	0.969	0.939	0.936	0.965	0.939
Weighted Avg.	**0.940**	**0.006**	**0.941**	**0.940**	**0.934**	**0.988**	**0.957**

Table 3. Results on the voice dataset

Class	TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area	PRC Area
Child	0.685	0.104	0.766	0.723	0.599	0.854	0.741
Female	0.778	0.125	0.757	0.767	0.648	0.856	0.728
Male	0.892	0.092	0.829	0.860	0.786	0.936	0.850
Weighted Avg.	**0.785**	**0.107**	**0.784**	**0.783**	**0.678**	**0.882**	**0.773**

Tables 4–6 present the confusion matrices obtained for the phishing, pollen, and voice datasets, respectively. In Table 4, the phishing identification results show a high number of correctly classified instances for both the positive and negative classes, indicating strong binary classification performance. Table 5 presents the confusion matrix for the pollen dataset, where most instances were correctly classified across the twelve categories, particularly for classes such as Seven and Ten, which achieved perfect classification. Some misclassifications were observed mainly between classes with similar characteristics, such as Two and Nine. Finally, Table 6 summarizes the results for voice identification. The male class obtained the highest number of correctly classified instances, while the child and female classes exhibited a greater number of misclassifications.

Table 4. Confusion matrix for phishing identification

	Predicted class	
Class	Positive	Negative
Positive	27947	2550
Negative	5466	51246

Table 5. Confusion matrix for pollen identification

	Predicted class											
Class	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Ten	Eleven	Twelve
One	44	0	0	0	0	0	0	0	1	0	1	0
Two	0	42	0	0	0	0	0	0	3	0	0	1
Three	0	0	54	1	0	0	0	0	0	0	1	0
Four	0	2	2	31	0	0	0	0	2	0	0	0
Five	0	1	0	0	41	0	0	0	0	0	0	0
Six	0	0	0	0	0	55	0	0	0	1	0	0
Seven	0	0	0	0	0	0	32	0	0	0	0	0
Eight	0	0	0	0	0	0	0	51	0	1	0	0
Nine	0	7	0	3	0	0	0	0	68	0	2	0
Ten	0	0	0	0	0	0	0	0	0	64	0	0
Eleven	0	0	2	0	0	0	0	0	3	0	68	0
Twelve	0	0	0	0	0	0	2	0	0	1	0	31

Table 6. Confusion matrix for voice identification

	Predicted class		
Class	Child	Female	Male
Child	196	58	32
Female	43	224	21
Male	17	14	257

Table 7 summarizes the general classification results obtained for the evaluated datasets using the algorithms automatically selected through the AutoML approach implemented in Weka. Among the evaluated cases, the pollen dataset achieved the best overall performance, obtaining a precision of 0.941. In contrast, the voice dataset presented the lowest performance metrics, achieving a precision of 0.784.

Table 7. General classification results

Classification	**Selected algorithm by AutoML**	**TP Rate**	**FP Rate**	**Precision**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**
Phishing	Random Forest	0.908	0.088	0.912	0.909	0.804	0.957	0.947
Pollen	Support Vector Machine	0.940	0.006	0.941	0.940	0.934	0.988	0.957
Voice	Random Forest	0.785	0.107	0.784	0.783	0.678	0.882	0.773

4 Discussion

The results obtained demonstrate the applicability of the SEMMA methodology across heterogeneous machine learning tasks involving tabular, image, and audio data. However, the observed performance should not be attributed exclusively to the SEMMA methodology, since the results also depend on the characteristics of the selected algorithms and datasets. In this study, AutoML selected Random Forest for the phishing and voice datasets, and Support Vector Machine for the pollen dataset, indicating that different models may be more suitable depending on the nature of the data.

For the phishing dataset (see Tables 2 and 4), the model achieved strong overall performance, obtaining a weighted average F-Measure of 0.909 and a ROC Area of 0.957. The confusion matrix also shows a high number of correctly classified instances for both positive and negative classes. Nevertheless, 2550 phishing instances were still misclassified as negative, which may be associated with overlapping characteristics between legitimate and phishing samples.

Regarding the pollen dataset (see Tables 3 and 5), the best overall performance was achieved among the evaluated datasets, with a weighted average F-Measure of 0.940 and ROC Area of 0.988. Several classes, including Seven and Ten, achieved near-perfect classification results in the confusion matrix. However, some confusion remained between classes such as Two and Nine, as well as Four and Nine, suggesting similarities in their extracted features.

For the voice dataset (see Tables 4 and 6), the results were comparatively lower than those obtained for the phishing and pollen datasets, with a weighted average F-Measure of 0.783 and ROC Area of 0.882. The confusion matrix reveals that the male class obtained the highest classification performance, while greater confusion occurred between the child and female classes. This behavior may be explained by similarities in certain acoustic patterns and variability in voice characteristics across speakers. Despite these challenges, the obtained results demonstrate the feasibility of using acoustic and statistical audio features for speaker identification tasks.

Overall, the comparative results across the three case studies suggest that the SEMMA methodology provides a structured framework for data preprocessing, modeling, and evaluation in heterogeneous machine learning applications. At the same time, the final predictive performance is strongly influenced by the selected classification algorithm, the dataset characteristics, and the complexity of the classification task.

5 Conclusions and Future Work

This study evaluated the applicability of the SEMMA methodology in phishing detection, pollen classification, and voice identification using tabular, image-derived, and audio datasets, respectively. The results demonstrate that SEMMA provides a structured framework for classification across different data analysis scenarios.

Additionally, experimental results showed that predictive performance varied depending on the characteristics of each dataset and the algorithm selected by AutoML. Random Forest performed well on the phishing and voice datasets, while Support Vector Machine achieved the best performance on the pollen dataset. Among the cases evaluated, the pollen dataset achieved the highest overall classification performance. In contrast, the voice dataset presented the most challenging task due to similarities in acoustic patterns across classes.

The main contributions of this study are summarized as follows:

1. The applicability of the SEMMA methodology was evaluated using heterogeneous datasets involving tabular, image-derived, and audio data.
2. An AutoML approach implemented via Auto-Weka was integrated to automate model selection and optimization.
3. Random Forest and Support Vector Machine were identified as the best-performing algorithms according to dataset characteristics and classification complexity.
4. A reproducible workflow combining SEMMA and AutoML for heterogeneous machine learning applications was provided.

Future work includes evaluating additional machine learning methodologies, integrating explainable artificial intelligence techniques to improve interpretability, and validating the proposed framework using larger and more diverse datasets. Future studies may also explore advanced visualization strategies and multimodal data fusion techniques to enhance classification performance and support decision-making processes.

References

Ahmad, Z., Yaacob, S., Ibrahim, R., & Wan Fakhruddin, W. F. (2022). The review for visual analytics methodology. In *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)* (pp. 1–10). IEEE. https://doi.org/10.1109/HORA55278.2022.9800100

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: A parallel overview. In H. Weghorn & A. P. Abraham (Eds.), *Proceedings of the IADIS European Conference on Data Mining 2008* (pp. 182–185). IADIS Press.

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing* (pp. 1–4). Springer. https://doi.org/10.1007/978-3-642-00296-0_5

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cios, K. J., Pedrycz, W., & Swiniarski, R. W. (1998). Machine learning. In *Data mining methods for knowledge discovery* (pp. 229–308). Springer.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

Costa, C. J., & Aparicio, J. T. (2020). POST-DS: A methodology to boost data science. In *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1–6). IEEE. https://doi.org/10.23919/CISTI49556.2020.9140932

De Ville, B. (2013). Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics, 5*(6), 448–455. https://doi.org/10.1002/wics.1278

De-La-Hoz-Correa, E., Mendoza-Palechor, F. E., De-La-Hoz-Manotas, A., Morales-Ortega, R. C., & Sánchez Hernández, B. A. (2019). Obesity level estimation software based on decision trees. *Journal of Computer Science, 15*(1), 67–77. https://doi.org/10.3844/jcssp.2019.67.77

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science, 14*(2), 241–258. https://doi.org/10.1007/s11704-019-8208-z

Duville, M. M., Alonso-Valerdi, L. M., & Ibarra-Zarate, D. I. (2021). The Mexican Emotional Speech Database (MESD): Elaboration and assessment based on machine learning. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 1644–1647). IEEE. https://doi.org/10.1109/EMBC46164.2021.9629934

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37–54. https://doi.org/10.1609/aimag.v17i3.1230

Fern, A., & Givan, R. (2003). Online ensemble learning: An empirical study. *Machine Learning, 53*(1–2), 71–109. https://doi.org/10.1023/A:1025619426553

Gama, J. (2010). *Knowledge discovery from data streams*. Chapman & Hall/CRC.

Hansun, S., Suryadibrata, A., Nurhasanah, R., & Fitra, J. (2022). Tweets sentiment on PPKM policy as a COVID-19 response in Indonesia. *Indian Journal of Computer Science and Engineering, 13*(1), 51–58.

Holzinger, A. (2013). Human-computer interaction and knowledge discovery (HCI-KDD): What is the benefit of bringing those two fields to work together? In A. Cuzzocrea, C. Kittl, D. E. Simos, E. Weippl, & L. Xu (Eds.), *Availability, reliability, and security in information systems and HCI* (Lecture Notes in Computer Science, Vol. 8127, pp. 319–328). Springer. https://doi.org/10.1007/978-3-642-40511-2_22

Jacob, D., & Henriques, R. (2023). Educational data mining to predict bachelors students' success. *Emerging Science Journal, 7*, 159–171. https://doi.org/10.28991/ESJ-2023-SIED2-013

Jadrić, M., Garača, Ž., & Čukušić, M. (2010). Student dropout analysis with application of data mining methods. *Management: Journal of Contemporary Management Issues, 15*(1), 31–46.

Li, G., Tan, J., & Chaudhry, S. S. (2019). Industry 4.0 and big data innovations. *Enterprise Information Systems, 13*(2), 145–147. https://doi.org/10.1080/17517575.2018.1554190

Lin, F. Y., & McClean, S. (2001). A data mining approach to the prediction of corporate failure. *Knowledge-Based Systems, 14*(3–4), 189–195. https://doi.org/10.1016/S0950-7051(01)00096-X

López-Torres, S., López-Torres, H., Rocha-Rocha, J., Butt, S. A., Tariq, M. I., Collazos-Morales, C., & Piñeres-Espitia, G. (2020). IoT monitoring of water consumption for irrigation systems using SEMMA methodology. In U. S. Tiwary & S. Chaudhury (Eds.), *Intelligent human computer interaction* (pp. 222–234). Springer. https://doi.org/10.1007/978-3-030-44689-5_20

Mohd Selamat, S. A., Prakoonwit, S., Sahandi, R., Khan, W., & Ramachandran, M. (2018). Big data analytics—A review of data-mining models for small and medium enterprises in the transportation sector. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(3), Article e1238. https://doi.org/10.1002/widm.1238

Palacios, H. J. G., Toledo, R. A. J., Pantoja, G. A. H., & Navarro, Á. A. M. (2017). A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change. *Advances in Science, Technology and Engineering Systems Journal, 2*(3), 598–604. https://doi.org/10.25046/aj020376

Panov, P., & Džeroski, S. (2007). Combining bagging and random subspaces to create better ensembles. In M. R. Berthold, J. Shawe-Taylor, & N. Lavrač (Eds.), *Advances in intelligent data analysis VII*

(Lecture Notes in Computer Science, Vol. 4723, pp. 118–129). Springer. https://doi.org/10.1007/978-3-540-74825-0_11

Rogalewicz, M., & Sika, R. (2016). Methodologies of knowledge discovery from data and data mining methods in mechanical engineering. *Management and Production Engineering Review, 7*(4), 97–108. https://doi.org/10.1515/mper-2016-0040

Sarker, O., Jayatilaka, A., Haggag, S., Liu, C., & Babar, M. A. (2024). A multi-vocal literature review on challenges and critical success factors of phishing education, training and awareness. *Journal of Systems and Software, 208*, Article 111899. https://doi.org/10.1016/j.jss.2023.111899

SAS Institute. (2003). *Data mining using SAS Enterprise Miner: A case study approach*.

Sazonau, V. (2012). *Implementation and evaluation of a random forest machine learning algorithm* [Unpublished report]. University of Manchester.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press. https://doi.org/10.1017/CBO9781107298019

Tello-Mijares, S., & Flores, F. (2016). A novel method for the separation of overlapping pollen species for automated detection and classification. *Computational and Mathematical Methods in Medicine, 2016*, Article 5689346. https://doi.org/10.1155/2016/5689346

Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013). Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 847–855). Association for Computing Machinery. https://doi.org/10.1145/2487575.2487629

Vistro, D. M., Rasheed, F., & David, L. G. (2019). The cricket winner prediction with application of machine learning and data analytics. *International Journal of Scientific & Technology Research, 8*(9), 985–990.

Vrbančič, G., Fister, I., Jr., & Podgorelec, V. (2020). Datasets for phishing websites detection. *Data in Brief, 33*, Article 106438. https://doi.org/10.1016/j.dib.2020.106438

Zavaleta-Sánchez, E., Domínguez-Sánchez, G., Loeza-Mejía, C.-I., & Sánchez-DelaCruz, E. (2024). Comparative study of KDD and CRISP-DM methodologies for phishing identification. In X.-S. Yang, S. Sherratt, N. Dey, & A. Joshi (Eds.), *Proceedings of Ninth International Congress on Information and Communication Technology* (Lecture Notes in Networks and Systems, Vol. 1013, pp. 317–330). Springer. https://doi.org/10.1007/978-981-97-3559-4_25

Zurada, J., & Karwowski, W. (2011). Knowledge discovery through experiential learning from business and other contemporary data sources: A review and reappraisal. *Information Systems Management, 28*(3), 258–274. https://doi.org/10.1080/10580530.2010.493846