# CIGI QUALITA MOSIM 2023

# Profiling the Applicability of Supervised Machine Learning in Supply Chain Information Systems: An Agglomerative Clustering Approach

SAMIA CHEHBI GAMOURA [1]

[1] EM Strasbourg Business School, Strasbourg University, HuManiS (UR 7308),
61 Av. de la Forêt-Noire, 67000 Strasbourg, France


***Abstract* – In recent years, Predictive Data Analytics (PDA) integrated with Supply Chain Information Systems (SCIS) has been the focus of considerable work to enable companies to make better decisions and remain competitive. Supervised Machine Learning (SML) approaches are becoming the main lever to smooth and facilitate this integration in this challenging environment. These approaches lead academics and industry to abandon procedural development and begin to think about undertaking them inductively by learning from the input data. However, SCISs have specific considerations that strongly affect the effectiveness of these learning models. Indeed, practitioners do not have some "recipe" for choosing a specific SML algorithm for a given problem in SCIS. They have to go through tedious phases and often depend on IT providers to do so. For this reason, the applicability of SML in SCIS is today an emerging challenge for scientists and practitioners, while the scientific literature is still in its early stages. This paper attempts to fill this gap by proposing a novel profiling approach for the applicability of SML in SCIS, including a comprehensive dual taxonomy with a Hierarchical Agglomerative Clustering algorithm (HAC). The profiling approach can help researchers and industrialists in the early selection of algorithms for their integration projects and thus avoid failure rates and expensive investments**

***Keywords* – Artificial Intelligence, Supervised Machine Learning, Supply Chain Information System, profiling, applicability.**

## INTRODUCTION

In the early 2000s, highly digitized systems started incorporating Predictive Data Analytics (PDA) to enhance decision-making and competitiveness in Supply Chains (SC) [Schoenherr et Speier-Pero, 2015]. One of the techniques used is Supervised Machine Learning (SML), a subcategory of Machine Learning (ML) and a branch of Artificial Intelligence (AI).

Supervised Machine Learning (SML) techniques have gained popularity with the continued growth of Supply Chain Information Systems (SCIS) [Mahraz et al., 2022], as they can learn from the environment and develop powerful models. These approaches effectively tackle complex decision problems where traditional methods fall short. For example, authors in [Harikrishnakumar et al., 2019] proposed SML algorithms including the Support Vector Machines (SVM), Logistic Regression (LRE), K-Nearest Neighbors (KNN), and Naïve Bayes Classifiers (NBC) algorithms. They used these models to classify various suppliers into four categories (excellent, good, satisfactory, and unsatisfactory) in the supplier assessment Multi-Criteria Decision-Making (MCDM) approach. Others like [Abbasi et al., 2020] proposed predicting solutions to deal with the issue of solving large optimization problems by using predictive algorithm models for decision makers. They propose predicting the optimal value of actionable decision variables.

Both industry and academia recognize SML as an opportunity to address management challenges by supporting decisions processes based on historical data models, outperforming the outdated techniques [Ni et al., 2020]. Examples include the use of Support Vector Machines in Enterprise Resource Planning (ERP) platforms [Chou et al., 2012] and Decision Trees (DTR) in Customer Relationship Management (CRM) systems [Chen et al., 2021].

Despite the growing adoption of SML techniques, several challenges and concerns have arisen regarding their use and applicability [Cavalcante et al., 2019]. The effectiveness of the resulting models heavily depends on the availability and quality of input data. There is no foolproof method to determine if a learning algorithm is appropriate, especially in predictive tasks that largely depend on data and system configuration [Cavalcante et al., 2019]. These challenges have made the applicability of SML in SCIS a crucial research problem for both industry and academia. However, academic research on this topic is still in its infancy. Hence, the need for systematic research in this field has become an urgent requirement.

This systemic investigation can be approached by comparing the applications of these algorithms using a criteria grid. However, the sheer number of algorithms, including meta-heuristics, heuristics, and their variants, would make a comprehensive comparison impractical in one study. A systematic and holistic approach is necessary to conduct such a research. To the best of our knowledge, such an approach has not been proposed in the existing scientific literature.

The main contribution of this paper is to fill this gap by introducing a holistic analytical framework to assess the applicability of SML algorithms in SCIS and provide a tool to guide practitioners and researchers.

This paper does not specifically focus on prescriptive analytics that aim to facilitate “actionable decision-making” [Lepenioti et

al., 2020] in accordance with the Gartner® Analytics Ascendancy Model (GAAM) [Eriksson et al., 2020], which focuses on "*How we can make things happen*". Although prescriptive analytics typically rely on predictive models, such as those utilizing predictive techniques as proposed in [Salah et Srinivas, 2022] and [Pessach et al., 2020], our present paper instead focuses on Supervised Machine Learning models in general. These models are geared towards predicting "*what will happen*" (as defined by the GAAM model [Eriksson et al., 2020], which can be used for both prescriptive and predictive analytics purposes as mentioned in [Bertsimas et Van Parys, 2022]. In other words, we do not primarily concentrate on foresighting analytics, but rather on insighting analytics, which serve as a key step in the decision-making process within organizations [den Hertog et Postek, 2016].

The article is structured as follows: Section 1 provides a literature review and background on SCIS and SML, identifies research gaps, and presents the problems in this area. Section 3 outlines our proposed methodology for addressing these gaps. Section 4 presents the main results. Finally, Section 5 concludes the paper with a summary, the limitations of our proposed solution, and suggestions for future work.

## 1 BACKGROUND AND RESEARCH GAP

The Supply Chain Information Systems (SCIS) comprise a group of interconnected software modules that work together to process, store, and control business operations. They convert data into valuable information across one or multiple Supply Chains (SC). The output of SCIS serves in the flow of information related to various management functions such as production, supply, marketing, sales, warehousing, accounting and strives to continuously improve performance [Akbari et Do, 2021]. The main drivers of this improvement can be reduced costs, enhanced customer services, and faster distribution, shipping, and delivery. These systems include Electronic Data Interchange (EDI), ERP platforms, Advanced Planning Systems (APS), Warehouse Management Systems (WMS), and many other supplementary systems.

With the advent of Big Data platforms, IoT, and Industry 4.0-related infrastructures that generate massive amounts of data, SCISs face new challenges that require more advanced analytics than conventional methods [Khan, 2013]. This allows for greater elasticity, adaptability, agility, and capabilities in prediction, real-time processing, optimization, and accuracy. Hence, SCISs have undergone several transformations over time, starting from theory-driven, practice-driven, and IT-driven, and now, we are in the data-driven era, where data is allowed to express itself.

Supervised Machine Learning (SML) is a subclass of Machine Learning techniques that operate on input-output pairs using "labeled" data, and the model can predict output classes based on that. Therefore, SML algorithms are also referred to as "induction classification algorithms." In other words, SML approaches can build models by mapping the relationship between the observed data features (input) and the labeled data (output) [Wuest et al., 2014]. Technically, the SML algorithm automatically generates a model by mapping the association between the Descriptive Data Features (DDF) (also known as "predictors") and the Targeted Output Variables (TOV). The TOV can be any generated values, either a discrete set of classes or a continuous range of values. After generating the model, new predictions can be made for newly observed data (Figure 1).

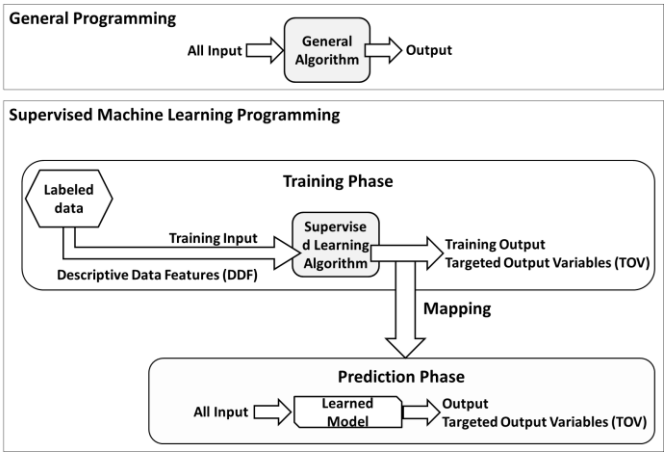



**Figure 1. Supervised ML algorithm vs. general algorithm**

The main challenges in integrating SML into Supply Chain Information Systems (SCIS) are related to the applicability and selection of algorithms. Practitioners face the challenge of choosing the appropriate algorithm for a specific business problem, as no single algorithm fits all cases, as demonstrated by the "No Free Lunch Theorems (NFLT)" [Wolpert et Macready, 1997]. Furthermore, IT developers do not always guarantee business value in SML-based supply chain systems [Reis et al., 2020]. The "curse of dimensionality" [Xu et al., 2020] can also impact the robustness of SML algorithms, as they can be sensitive to the size of the dataset. Additionally, the implementation and training phases of SML require significant time and investment, making pre-selection of algorithms crucial for successful application. Despite their growing use, the development of successful SML applications still requires a significant amount of “black art” and investment.

## 2 RESEARCH METHODOLOGY

Our review of the existing literature revealed two distinct parts in the literature review studies: (1) Technique-focused LR studies, where the authors only focused on a specific class of algorithms and did not consider all SML approaches, as seen in [Rostami et al., 2015]. (2) Domain-focused LR studies, where the authors analyzed one or more specific domains in management but lacked a comprehensive scope, as seen in [Akbari et Do, 2021]. Given these challenges and the gaps in the research, our paper aims to provide a clear understanding of the applicability profiles of SML in SCIS to guide current and future research.

Due to the complexity and diversity of SML techniques and SCIS categories, we chose a two-stage methodology for this study, starting with a primary taxonomy, followed by a detailed examination and profiling in the second stage.

### *2.1 Stage of Taxonomification*

Our approach is based on a dual taxonomy, consisting of: (1) a taxonomy for Supply Chain Information Systems (SCIS) based on business functions, and (2) a taxonomy for Supervised Machine Learning (SML) techniques based on current business applications. The next two sub-sections will provide further detail on each of these proposed taxonomies.

#### 2.1.1 Proposed SCIS Taxonomy

The examination of the intersection of research on SCIS reveals the use of two commonly recognized classifications: (1) The levels-oriented classification of Ivanov (2010) (SCIS-C1),

which is based on four levels: operational-level systems (OLS), management-level systems (MLS), knowledge-level systems (KLS), and strategic-level systems (SLS). (2) The functions-oriented classification of Bucher and Winter (2010) (SCIS-C2), which is based on four functions: sales and marketing function (SMF), finance and accounting function (FAF), manufacturing, logistics, and production function (MPF), and human resources and workforce function (HRW).

Despite the widespread use of these classifications in research, some SCISs involve cross-functional operations that are challenging to categorize separately, such as business process management (BPM). As a result, adopting these classifications may lead to siloed functions, which would be restrictive and exclude some functions. To address this issue, we propose a third combined classification inspired by the typologies in Huemann (2010) and Hwang et al. (2015). The four classes of this proposed taxonomy (named "SCIS-$C^3$") are as follows:

1. Transaction Information Systems (TIS),
2. Automation, Knowledge, & Engineering systems (AKE),
3. Decision Support Systems (DSS),
4. Macro Planning Systems (MPS).

#### 2.1.2 Proposed SML Taxonomy

Research indicates three main categories of SML in SCIS: (1) Theory-oriented classification by Taiwo [Taiwo, 2010] (SML-C1), which consists of twelve classes based on the ground theory of classification, including Linear Classifiers (LCL), Logistic Regression, Naive Perceptron Classifiers (NPC), Support Vector Machines (SVM), Bayes Classifiers (BCL), Polynomial Classifiers (PCL), Boosting (BOO), Decision Trees (DTR), Random Forest (RFO), Artificial Neural Networks (ANN), K-Means Clustering (KMC), K-Nearest Neighbor Classifiers, and Bayesian Networks (BNE). (2) Mechanism-oriented classification by Kotsiantis et al. [Kotsiantis et al., 2007] (SML-C2), which categorizes SML into five classes based on operational mechanism: Logic-Based Techniques (LBT), Perceptron-Based Techniques (PBT), Statistics-Based Techniques (SBT), and Support Vector Machines (SVM). (3) Problem-oriented classification, the most commonly used in research [Rodrigues et al., 2017] (SML-C3), which divides SML into two types based on problem nature: Classification Algorithms (CLA) and Regression Algorithms (REA). The Classification Algorithms class predicts the group to which the data belongs and its output consists of only discrete unordered values.

SCISs are highly dynamic, uncertain, and contain huge unstructured data with duplications, redundancies, and noise, as well as silos of truncated business datasets [Akbari et Do, 2021]. Unlike ISs in other fields, such as healthcare, where medical data is usually structured. For these reasons, we suggest that the focus should be on the predictor features (DDF) rather than just the theory-oriented, target-oriented, or problem-oriented classification. Hence, the classification we propose is DDF-oriented and is adapted from classifications by [Kelleher et al., 2020] and [Xu et al., 2007]. It consists of four classes as follows:

1. Information-Based Learning Approaches (INF): This category's fundamental basis is Shannon's theory presented in 1948 by Claude Shannon [Rodrigues et al., 2018]. The key idea is to extract a specific measurement from the DDF through the information contents.
2. Error-Based Learning Approaches (ERR): Error-based approaches are mathematical models that are founded on the idea of getting performance through error minimization in the training phase [Kelleher et al., 2020].
3. Similarity-Based Learning Approaches (SIM): The key idea in these approaches is that the best way of predicting the future is to compare and find similarities in the past [Chen et al., 2009]. That leads to determining similarities in the defining features of DDF.
4. Probability-Based Learning Approaches (PRO): They are based on the Bayes theorem [Kelleher et al., 2020]. The idea is that the future is a random event based on relative frequencies with the calculation of conditional probabilities based on the present and the past.

Table 1 below presents the key publications reviewed during the process, including the identification of the relevant SCOR processes.

### *2.1 Stage of Review Process*

#### 2.1.1 Dataset Extraction

We used the Harzing Publish or Perish® [Harzing.com, 2022] tool over the selected period of publications from 2010 to 2018. Due to the significant impact of COVID-19 on publications after 2018, we decided to end the extraction at that date. The dataset was extracted from various academic search engines including Crossref, Microsoft Academic, Google Scholar, and Scopus and obtained from databases such as Springer, Wiley, Elsevier, Taylor & Francis, IEEE, Emerald, and Inderscience. The initial dataset yielded 1,000 papers in a CSV list with columns such as title, authors, citations, publication year, publisher, and source URL. We then filtered the sorted dataset by removing unwanted columns, such as ISSN, GSRank, sources, and links, and kept only the relevant columns for our analysis, including authors, citations, publication year, title, journal name, and URL. We excluded all papers that were not in English, editorial pieces, reports, theses, white papers, conference or seminar papers, patents, and papers on irrelevant topics. This process reduced the initial raw list of 1,000 to 957 papers.

#### 2.1.2 Dataset Preparation

The list of papers was divided chronologically into three groups to facilitate analysis: group 1 (2010-2013), group 2 (2014-2016), and group 3 (2017-2018). During the first scan, we briefly reviewed the abstracts, introductions, and conclusions of each of the 957 selected papers. In the second scan, the papers were systematically divided into three categories based on the type of publication: original research, surveys, and review papers. The results showed that the usage of LR studies was negligible compared to original research and surveys. Finally, the cleansing step reduced the list to 830 mixed original research papers that were reclassified for analysis, while LR and survey papers were excluded as they were not relevant in determining applicability profiles.

**Table 1. Illustrative examples of some relevant publications with proposed classifications vs. existing classifications**

| Research publication | SCOR Processes * | SCIS Business function(s) | SCIS classifications | | | Used SML algorithms | SML classifications | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | SCIS-C[1] | SCIS-C[2] | SCIS-C[3] | | SML-C[1] | SML-C[2] | SML-C[3] | SML-C[4] |
| [Ullah et al., 2019] | D | Churn prediction | MLS | SMF | TIS | Random Forest | RFO | LBT | REA | INF |
| [Anbazhagan et Kumarappan, 2012] | D | Markets forecasting | KLS | SMF | AKE | Recurrent Neural Networks | NNE | PBT | CLA | ERR |
| [Hua, 2011] | D | Customer management | MLS | SMF | TIS | Naive Bayesian classifier | BCL | SBT | CLA | PRO |
| [Kwon, 2017] | D | Transportation | OLS | MPF | TIS | Artificial Neural Network | NNE | PBT | CLA | ERR |
| [De Paula et al., 2019] | P | Profit scoring | OLS | FAF | TIS | Logistic Regression | LRE | SBT | REA | |
| [Ryu et al., 2020] | S | Purchase forecasting | MLS | MPF | DDS | Long-short Time Memory | NNE | PBT | CLA | ERR |
| [Zhang et al., 2017] | P | Project Management | SLS | SMF | DDS | Lasso Regression | LRE | PBT | CLA | ERR |
| [Ganji et Mannem, 2012] | D | Fraud Detection | OLS | FAF | TIS | K-Nearest Neighbor | KNN | SBT | CLA | SIM |
| [Gao et Fan, 2021] | M | Customer experience | OLS | MPF | TIS | Polynomial regression | PCL | SBT | REA | ERR |
| [Rahimi, 2017] | D | Customer management | MLS | MPF | DDS | Linear Regression | LRE | SBT | REA | ERR |
| [Ju et al.,2017] | M | Product processes | MLS | MPF | DDS | Least-angle regression OLS | LRE | SBT | REA | ERR |
| [Chou et al., 2012] | P | Forecasting in ERP | SMF | AKE | SMF | Support Vector Machine | SVM | SVM | CLA | ERR |
| [Chen et al., 2021] | P | Customer management | MLS | MPF | DDS | Decision Trees | DTR | LBT | REA | INF |
| [Fei et al., 2017] | D | Customer churn | MLS | SMF | DSS | Naive Bayes Classifier | BCL | SBT | CLA | PRO |
| [Lockamy III, 2011] | S | Suppliers management | MLS | MPF | DSS | Bayesian Networks | BNE | SBT | REA | PRO |
| [Al-Dmour et Al-Dmour, 2018] | P | Performance prediction | MLS | FAF | DSS | Multiple Linear Regression | LCL | SBT | REA | ERR |
| [Von Kirby et al., 2017] | D | Sales classification | KLS | SMF | AKE | AdaBoost Algorithm | BOO | LBT | CLA | INF |
| [Kirori, 2011] | P | Banking and funding | KLS | FAF | AKE | LogitBoost Algorithm | BOO | LBT | CLA | INF |

** SCOR processes: P: Plan, S: Source, M: Make, D: Deliver, R: return*

In order to construct the Dataset, we first conducted an extensive scan of the resulted 830 papers, analyzing their characteristics and matching them with the matrix formed by the SCIS proposed taxonomy and the SML proposed taxonomy (as described in sections 3.1.1 and 3.1.2). Next, we computed the number of papers that corresponded to each cell in the matrix in order to normalize the data. As a result, we obtained a matrix that dynamically maps the relationship between the SML approaches and the SCIS classes, as presented in Table 2.

**Table 2. Simplified illustration of the Dataset (Matrix SCIS x SML)**

| Classes in Taxonomis (SCIS X SML) | | | | Automation, knowledge, and Engineering (AKE) | Decision Support Systems (DSS) | Macro Planning Systems (MPS) | Transaction Information Systems (TIS) |
|---|---|---|---|---|---|---|---|
| Information-Based Learning Approaches | Ensemble | Bagging | AdaBoost Algorithm (ABA) | 0.0000% | 0.3425% | 0.0000% | 1.0695% |
| Error-Based Learning Approaches (EBLA) | Supervised | Classification & prediction | Artificial Neural Network (ANN) | 3.7901% | 4.1096% | 4.0404% | 5.8824% |
| Information-Based Learning Approaches | Unsupervised | Association Rule | Apriori Algorithm (AAL) | 0.5831% | 1.0274% | 0.0000% | 0.5348% |
| Probability-Based Learning Approaches | Supervised | Classification & prediction | Bayesian Networks (B | 1.4577% | 1.0274% | 3.0303% | 2.6738% |
| Similarity-Based Learning Approaches | Supervised | Classification & Prediction | Case-Bases Reasoning (CBR) | 0.5831% | 2.0548% | 2.0202% | 2.6738% |
| Information-Based Learning Approaches | Supervised | Classification & Prediction | Conditional Random Fields (CRF) | 1.7493% | 2.0548% | 3.0303% | 1.0695% |
| ... | ... | ... | ... | ...% | ...% | ...% | ...% |

### *2.2 Stage of Analytics*

The profiling techniques aim to identify distinct "characterizable" shapes within sets of objects, serving as a fundamental method to uncover hidden patterns, understand information distributions, and identify dissimilarities in datasets [Ayanso et Yoogalingam, 2009]. Clustering techniques have been widely used in profiling, including in areas of management such as customer behavior segmentation [Tsourgiannis et Valsamidis, 2021] and supplier profiling [Visani et Boccali, 2020], among others.

The Hierarchical Agglomerative Clustering (HAC) approach is a widely used bottom-up clustering algorithm that is effective in classification, pattern retrieval, and profiling, particularly with multi-dimensional variables. We applied the HAC algorithm to the preliminary phase dataset using the R scripting language, with the Ward distance metric. The Ward-D2 (AGNES with k=4) was calculated between the profiles (clusters) based on the squared Euclidean distance.

The Ward-D2 distance metric $D(C_1, C_2)_{Ward}$ is expressed as :

$$D(C_1, C_2)_{Ward} = \frac{(N_1 * N_2)}{(N_1 + N_2)} \times D(C_1, C_2)_{Euclidian} \quad (1)$$

Where:

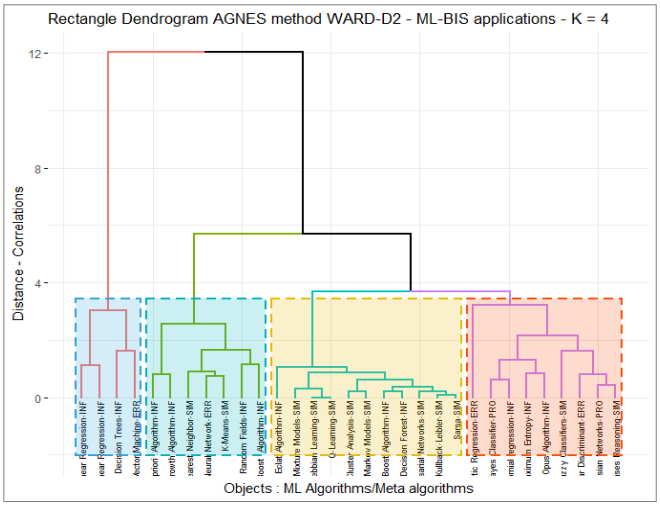
$N_1$ and $N_2$ are the number of observations in clusters C1 and C2, respectively,

$x_i \in X_i$ and $x_j \in X_j$ are the data in the clusters of C1 and C2 respectively, where each row represents a data point.

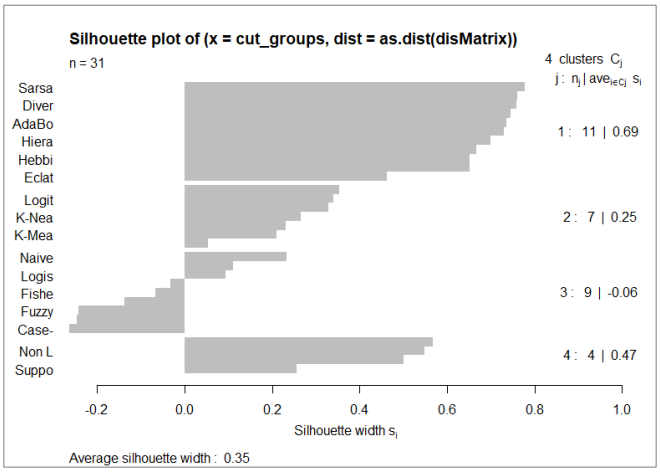
$D(C_1, C_2)_{Euclidian}$ is the squared Euclidean distance between the centroids of these clusters expressed as:

$$D(C_1, C_2)_{Euclidian} = \frac{1}{(N_1 \times N_2)} \times \sum_i^{N_1} \sum_j^{N_2} \|x_i - x_j\|^2 \quad (2)$$

The results of the HAC applied to our dataset are shown in Figure 2, Figure 3 and Table 3 below.




**(a)**




**(d)**

**Figure 2. HAC representations: (a) rectangular dendrogram, (b) Silhouette diagram (AGNES method WARD-D2, K = 4, developed with R®)**

2.2.1 Inclusion of SML classes in profiles (clusters)

Figure 2 depicts the rectangular and silhouette diagrams, which illustrate four clusters (1, 2, 3, and 4) comprising 11, 7, 9, and 4 classes of SML approaches, respectively. It can be observed that Cluster 1 is the most prolific, followed by Cluster 3, Cluster 2, and finally Cluster 4. Table 3 provides a complete description of the four clusters (profiles), which is complemented by Figure 3. These figures reveal that all the profiles (clusters) have adopted SML approaches to varying degrees and intensities.

Moreover, Table 3 indicates that the presence or absence of certain SML classes distinguishes and identifies the profiles (clusters), namely Information-based (INF), Error-based (ERR), Similarity-based (SIM), and Probability-based (PRO). For example, Profile 1 includes all classes, whereas Profile 2 comprises 3 (INF, ERR, SIM), Profile 3 includes 2 (INF, ERR), and Profile 4 contains only 1 (INF).

2.2.1 Inclusion of SCIS classes in profiles (clusters)

Figure 3 and Table 3 reveal the following: Cluster 1 includes 61 ML applications of 11 ML algorithms with a variation exponential trend of ($y = 0.4795e^{-0.387x}$, $R^2 \sim 0.26$). The first position is to hold by the knowledge-based applications (AKE) with the majority of more than half (54.10%), followed by moderate intensities in the transactional systems (TIS) (22.95%), then the decision-based platforms (DSS) with 18.03%. Only a few works (4.92%) are placed in the planning systems' last position (MPS). The profile of Cluster 2 covers 256 applications using 8 ML approaches with a variation of ($y = 0.3299e^{-0.207x}$, $R^2 = 0.08$) in the following ranking: the decision systems (DSS) take the first position modestly with 34.38%, followed narrowly by the knowledge systems (AKE) with 30.86%, then the transaction-based applications (TIS) with 30.08%, and lastly the macro-planning platforms (MPS) with few works (4.69%). The profile of Cluster 3 comprises 225 publications embedding 9 ML algorithms with an exponential trend of ($y = 0.5923e^{-0.391x}$, $R^2 = 0.8013$). The knowledge-based systems (AKE) are ranked in the first position with 39.56% neighboring the decision-based systems (DSS) with 32.89%, and the transaction-based applications (TIS) with 14.67%, and lastly, we find the macro-planning systems (MPS) with more than 12.-%. The profile corresponding to the Cluster 4 contains only 4 ML algorithms that have been used in more than 300 applications (~379) with an exponential trend of ($y = 0.5149e^{-0.321x}$, $R^2 = 0.7895$). These applications represent mainly the knowledge systems (AKE) but not in an overwhelming majority (38.-%), followed by the decision systems (DSS) with 31.40%. The remaining applications are then led by the transaction-based platforms (TIS) with 16.62% and the planning systems (MPS) with 14.51%.

All the profiles incorporate Information-based algorithms (INF), but at different frequencies. We attribute this observation to their efficiency and extensive practical experience in SC applications. This category of approaches has the advantage of being applied to multiple predictive models [Kelleher et al., 2020], which is crucial due to the lack of technical and AI-based business skills among practitioners. Notable examples are Decision Trees algorithms [Chen et al., 2021] and Random Forest algorithms [Ryu et al., 2020]. However, their main drawback is the selection of content metrics, which can be challenging, especially with noisy data.

Error-based algorithms (ERR) are included in three profiles (1, 2, and 3). Despite their proven and long-standing use in other fields, they are not universally applicable in SCIS as they require data-rich datasets for the training phase and continuous error correction. However, they are still preferred when these conditions are met, such as in Artificial Neural networks [Kwon, 2017]. As a result, several new variants of algorithms and hybrid heuristics are gradually emerging to overcome this limitation. An example is the combination of Fuzzy Logic and Genetic Algorithm to avoid local minima in ANN training [Azeem et Mohammad, 2015].

**Table 3. Numerical characteristics of the four clusters from HAC applied to the SML-SCIS mapping**

| | *SCIS Applications* | | | | | | *ML Approaches* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Totals* | | *AKE* | *DSS* | *MPS* | *TIS* | *Algorithms, Meta-algorithms* | *Totals* | *Information-based* | | *Error-based* | | *Similarity-based* | | *Probability-based* | |
| | | | % | % | % | % | | | # | % | # | % | # | % | # | % |
| *Cluster 1* | # 61 | *Max* | 33.33 | 27.27 | 33.33 | 21.43 | *AdaBoost, Divergence de Kullback-Leibler, Eclat, Generative Adversarial Networks, Hebbian Learning, Hierarchical Cluster Analysis, Hidden Markov Models, Mixture Models, Q-Learning, Random Decision Forest, Sarsa* | # 11 | 64 | 7.71% | 66 | 7.95% | 31 | 3.73% | 44 | 5.30% |
| | | *Average* | 9.09 | 9.09 | 9.09 | 9.09 | | | ⇧ INF | | ⇧ ERR | | ⇧ SIM | | ⇧ PRO | |
| | | *Min* | 0.00 | 0.00 | 0.00 | 0.00 | | | | | | | | | | |
| | | *All* | 54.10 | 18.03 | 4.92 | 22.95 | | | | | | | | | | |
| *Cluster 2* | # 256 | *Max* | 24.05 | 28.41 | 33.33 | 16.88 | *Artificial Neural Network, Apriori, Conditional Random Fields, FP Growth, K-Means, K-Nearest Neighbor, LogitBoost* | # 7 | 39 | 4.70% | 40 | 4.82% | 47 | 5.66% | 0 | 0.00% |
| | | *Average* | 14.29 | 14.29 | 14.29 | 14.29 | | | ⇧ INF | | ⇧ ERR | | ⇧ SIM | | ⇧ 000 | |
| | | *Min* | 2.53 | 2.27 | 0.00 | 11.69 | | | | | | | | | | |
| | | *All* | 30.86 | 34.38 | 4.69 | 30.08 | | | | | | | | | | |
| *Cluster 3* | # 225 | *Max* | 33.71 | 21.62 | 27.59 | 24.24 | *Bayesian Networks, Case-Bases Reasoning, Fisher Linear Discriminant, Fuzzy Classifiers, Logistic Regression, Maximum Entropy, Naive Bayes Classifier, Opus, Polynomial regression* | # 9 | 200 | 24.10% | 82 | 9.88% | 0 | 0.00% | 0 | 0.00% |
| | | *Average* | 11.11 | 11.11 | 11.11 | 11.11 | | | ⇧ INF | | ⇧ ERR | | ⇧ 000 | | ⇧ 000 | |
| | | *Min* | 0.00 | 1.35 | 0.00 | 0.00 | | | | | | | | | | |
| | | *All* | 39.56 | 32.89 | 12.89 | 14.67 | | | | | | | | | | |
| *Cluster 4* | # 379 | *Max* | 28.17 | 30.25 | 34.55 | 31.75 | *Decision Trees, Linear Regression, Non-Linear Regression, Support Vector Machine* | # 4 | 217 | 26.14% | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% |
| | | *Average* | 25.00 | 25.00 | 25.00 | 25.00 | | | ⇧ INF | | ⇧ 000 | | ⇧ 000 | | ⇧ 000 | |
| | | *Min* | 20.42 | 18.49 | 14.55 | 17.46 | | | | | | | | | | |
| | | *All* | 37.47 | 31.40 | 14.51 | 16.62 | | | | | | | | | | |

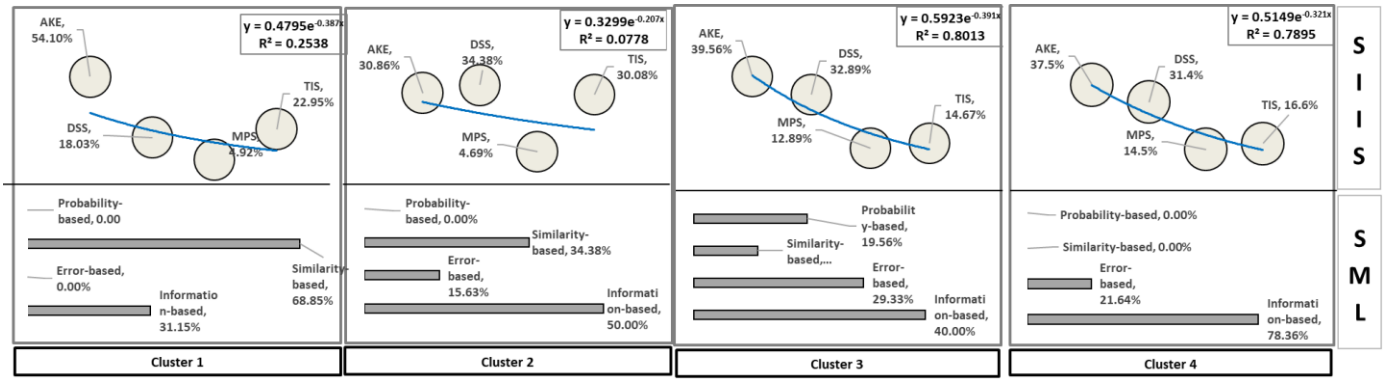



**Figure 3. Graphical characteristics of the four clusters resulted from HAC applied to the SML-SCIS**

Thirdly, the Similarity-based algorithms (SIM) are present in two profiles (1, 2). Firstly, they are suitable for situations requiring rapid scalability in business systems, such as the Nearest Neighbors (KNN) [Ganji et Mannem, 2012] in cases of significant data context requiring parallel processing. The main challenge with similarity-based models is their high memory usage, particularly when dealing with large-scale data where binary distances increase exponentially.

Lastly, Probability-based algorithms (PRO) are less frequently used in SCIS and constitute a small portion of the overall algorithms. Despite their proven efficiency in various disciplines due to their ability to train rapidly [Lockamy III, 2011], they are only present in profile 1. The advantage of probability-based approaches stems from their grounding in Bayesian theories, which are influential in various targeted feature classes. Therefore, they are better suited for SCIS, which demands robust real-time results.

## 3 INTERPRETATION AND DISCUSSION

Based on the above observations, the following are the distinctive characteristics of the profiles:

**1.** Data set density is crucial for SML algorithms in general. However, when IS interacts with SCM, this specific factor intensifies challenges such as noisy data, confidentiality, and security issues. In the profiles, the intensity of data availability was found to differentiate between data-rich and data-poor environments, which constitutes a defining characteristic. Examples of data-rich environments are significant data contexts connected to SCM, such as industrial-technological processes [Wamba et al., 2015]. On the other hand, cases of data-poor environments correspond more to emerging sectors of SCM where knowledge and records are still missing, such as fin-tech from emerging online banks [Bazarbash, 2019].

2. Scalability also defines distinctive features of profiles for SML algorithms. For instance, it is necessary when the dataset size may grow, and the SML model must dynamically adjust accordingly [Gupta et al., 2016]. We observed that the need for scalability (or not) identifies specific profiles. For example, some SML algorithms may struggle when scalability is applied as they fail when their inputs scale to larger datasets, such as RFID-based decision systems in distributed manufacturing [Guo et al., 2015]. However, cases of non-scalable systems characterize sectors where dataset dimensions are still stable.

3. Real-time SML in continuous transactional data and streaming analytics is a strict requirement in some sectors, such as RFID-enabled real-time logistics trajectory [Zhong et al., 2015]. However, not all SML models can remain robust and operate with responsiveness in real-time mode. Most of these algorithms experience latency when challenged with continuous flow processing.

## 4 CONCLUSION, LIMITATIONS, AND OPEN VIEWS

The focus of this paper is on the applicability of Supervised Machine Learning (SML) in Supply Chain Information Systems (SCIS). We present a comprehensive study of academic SML-based SCIS applications through a double-taxonomy and clustering-based profiling approach. The results of our analysis led to the identification of four distinct profiles. Our main findings indicate that the applicability of SML approaches is influenced by the data availability and maturity in SCIS, and these factors determine the four profiles. Secondly, the disparity in applying SML algorithms is not driven by organizational or managerial needs, but rather by the ability of the business system to provide the necessary data features, including data richness, scalability, and real-time processing.

The use of a Hierarchical Agglomerative Clustering algorithm for profiling constitutes an originality in our approach, as it is a suitable technique for analyzing complex datasets. However, this approach also has limitations, as clustering can be challenging when determining the number of clusters, and it can be sensitive to outliers. Additionally, the use of Harzing® for dataset extraction is limited to 1000 rows, despite its user-friendly interface.

Our future research direction is to develop an assessment model of applicability based on the four profiles. This model will enable the evaluation and comparison of SML applicability for specific business problems prior to development and training.

## 5 REFERENCES


Abbasi, B., Babaei, T., Hosseinifard, Z., Smith-Miles, K., & Dehghani, M. (2020). Predicting solutions of large-scale optimization problems via machine learning: A case study in blood supply chain management. *Computers & Operations Research*, 119:104941.

Akbari, M., & Do, T. (2021). A systematic review of machine learning in logistics and supply chain management: current trends and future directions. *Benchmarking: An International Journal*.

Al-Dmour, A. H., & Al-Dmour, R. H. (2018). Applying Multiple Linear Regression and Neural Network to Predict Business Performance Using the Reliability of Accounting Information System. *International Journal of Corporate Finance and Accounting (IJCFA)*, 5(2), 12-26.

Anbazhagan, S., & Kumarappan, N. (2012). Day-ahead deregulated electricity market price forecasting using recurrent neural network. *IEEE Systems Journal*, *7*(4), 866-872.

Ayanso, A., & Yoogalingam, R. (2009). Profiling retail web site functionalities and conversion rates: A cluster analysis. *International Journal of Electronic Commerce*, 14(1), 79-114.

Azeem, M. M., & Mohammad, A. (2015). An analysis of applications and possibilities of neural networks (fuzzy, logic and genetic algorithm) in finance and accounting. *Donnish Journal of Business and Finance Management Research*, 1(2), 9-18.

Bazarbash, M. (2019). Fintech in financial inclusion: machine learning applications in assessing credit risk. *International Monetary Fund*.

Bertsimas, D., & Van Parys, B. (2022). Bootstrap robust prescriptive analytics. *Mathematical Programming*, 195(1-2):39-78.

Bucher, T., & Winter, R. (2010). Taxonomy of business process management approaches. In B. S. Vom Brocke, *Handbook on Business Process Management 2: Strategic Alignment, Governance, People and Culture* (pp. 93-114). Springer, Berlin, Heidelberg: J. M. Rosemann (Ed.).

Cavalcante, I., Frazzon, E., Forcellini, F., & Ivanov, D. (2019). A supervised machine learning approach to data-driven simulation of resilient supplier selection in digital manufacturing. *International Journal of Information Management*, *49*, 86-9.

Chen, C., Geng, L., & Zhou, S. (2021). Design and implementation of bank CRM system based on decision tree algorithm. *Neural Computing and Applications*, *33*(14), 8237-8247.

Chen, Y., Garcia, E. K., Gupta, M. R., Rahimi, A., & Cazzanti, L. (2009). Similarity-based classification: Concepts and algorithms. *Journal of Machine Learning Research*, 10,747-776.

Chou, J. S., Cheng, M. Y., Wu, Y. W., & Wu, C. C. (2012). Forecasting enterprise resource planning software effort using evolutionary support vector machine inference model. *International Journal of Project Management*, 30(8), 967-977.

De Paula, D. A., Artes, R., Ayres, F., & Minardi, A. M. (2019). Estimating credit and profit scoring of a Brazilian credit union with logistic regression and machine-learning techniques. *RAUSP Management Journal*.

den Hertog, D., & Postek, K. (2016). Bridging the gap between predictive and prescriptive analytics-new optimization methodology needed. *Tilburg Univ, Tilburg, The Netherlands*.

Eriksson, T., Bigi, A., & Bonera, M. (2020). Think with me, or think for me? On the future role of artificial intelligence in marketing strategy formulation. *The TQM Journal*, 32(4): 795-814.

Fei, T., Shuan, L., Yan, L., Xiaoning, G., & King, S. (2017). Prediction on customer churn in the telecommunications sector using discretization and Naïve Bayes Classifier. *International Journal Advance Soft Computing Applications*, *9*(3), 23-34.

Ganji, V. R., & Mannem, S. N. (2012). Credit card fraud detection using anti-k nearest neighbor algorithm. *International Journal on Computer Science and Engineering*, 4(6), 1035-1039.

Gao, W., & Fan, H. (2021). Omni-channel customer experience (in) consistency and service success: a study based on polynomial regression analysis. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*(6), 1997-2013.

Guo, Z., Ngai, E., Yang, C., & Liang, X. (2015). An RFID-based intelligent decision support system architecture for production monitoring and scheduling in a distributed manufacturing environment. *International journal of production economics*, 159, 16-28.

Gupta, P., Sharma, A., & Jindal, R. (2016). Scalable machine-learning algorithms for big data analytics: a comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *6*(6), 194-214.

Harikrishnakumar, R., Dand, A., Nannapaneni, S., & Krishnan, K. (2019). Supervised machine learning approach for effective supplier classification. Dans *IEEE International Conference*

*On Machine Learning And Applications (ICMLA)* (pp. 240-245). IEEE.

Harzing.com. (2022, 01 01). *Harzing's Publish or Perish*. (Harzing.com) Consulté le 09 20, 2022, sur https://harzing.com/resources/publish-or-perish

Hua, G. (2011). Customer relationship management based on data mining technique—Naive Bayesian classifier. Dans *In proceedings of International Conference on E-Business and E-Government (ICEE)* (pp. 1-4). IEEE.

Huemann, M. (2010). Considering Human Resource Management when developing a project-oriented company: Case study of a telecommunication company. *International Journal of Project Management*, 28(4), 361-369.

Hwang, B. G., Zhao, X., & Ong, S. Y. (2015). Value management in Singaporean building projects: Implementation status, critical success factors, and risk factors. *Journal of management in engineering*, 31(6), 04014094.

Ivanov, D. (2010). An adaptive framework for aligning (re) planning decisions on supply chain strategy, design, tactics, and operations. . *International journal of production research*, 48(13), 3999-4017.

Ju, L., Zhou, J., & Zhang, X. (2017). June. Corundum production quality prediction based on support vector regression. Dans *12th IEEE Conference on Industrial Electronics and Applications (ICIEA)* (pp. 2028-2032). IEEE.

Kelleher, J. D., Mac Namee, B., & D'arcy, A. (2020). *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies.* New York, USA.: MIT press.

Khan, R. (2013). Business analytics and supply chain performance: An Empirical Perspective. *International Journal of Operations and Logistics Management, 2*(3), 43-56.

Kirori, Z. (2011). *A System for Credit Appraising-An application of the LogitBoost Algorithm.* Nairoby, Kenya: University of Nairoby, Doctoral dissertation.

Kotsiantis, S., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering, 160*(1), 3-24.

Kwon, H. B. (2017). Exploring the predictive potential of artificial neural networks in conjunction with DEA in railroad performance modeling. *International Journal of Production Economics*, 183, 159-170.

Lepenioti, K., Bousdekis, A., Apostolou, D., & Mentzas, G. (2020). Prescriptive analytics: Literature review and research challenges. *International Journal of Information Management*, 50:57-70.

Lockamy III, A. (2011). Benchmarking supplier risks using Bayesian networks. *Benchmarking: An International Journal*, 18(3), 409-427.

Mahraz, M., Benabbou, L., & Berrado, A. (2022). Machine Learning in Supply Chain Management: A Systematic Literature Review. *International Journal of Supply and Operations Management, 9*(4), 398-416.

Ni, D., Xiao, Z., & Lim, M. (2020). A systematic review of the research trends of machine learning in supply chain management. *International Journal of Machine Learning and Cybernetics, 11*(7), 1463-1482.

Pessach, D., Singer, G., Avrahami, D., Ben-Gal, H., Shmueli, E., & Ben-Gal, I. (2020). Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming. *Decision Support Systems*, 134:113290.

Rahimi, R. (2017). Organizational culture and customer relationship management: A simple linear regression analysis. *Journal of Hospitality Marketing & Management, 26*(4), 443-449.

Reis, C., Ruivo, P., Oliveira, T., & Faroleiro, P. (2020). Assessing the drivers of machine learning business value. *Journal of Business Research, 117*, 232-243.

Rodrigues, F., Lourenco, M., Ribeiro, B., & Pereira, F. (2017). Learning supervised topic models for classification and regression from crowds. *IEEE transactions on pattern analysis and machine intelligence, 39*(12), 2409-2422.

Rodrigues, M., Bolcskei, H., Draper, S., Eldar, Y., & Tan, V. (2018). Introduction to the Issue on Information-Theoretic Methods in Data Acquisition, Analysis, and Processing. *IEEE Journal of Selected Topics in Signal Processing*, 12(5), 821-824.

Rostami, H., Dantan, J., & Homri, L. (2015). Review of data mining applications for quality assessment in manufacturing industry: support vector machines. *International Journal of Metrology and Quality Engineering, 6*(4), 401.

BIBLIOGRAPHY Ryu, G., Nasridinov, A., Rah, H., & Yoo, K. (2020). Forecasts of the amount purchase pork meat by using structured and unstructured big data . *Agriculture*, 10(1): 21.

Salah, H., & Srinivas, S. (2022). Predict, then schedule: Prescriptive analytics approach for machine learning-enabled sequential clinical scheduling. *Computers & Industrial Engineering*, 169:108270.

Schoenherr, T., & Speier-Pero, C. (2015). Data science, predictive analytics, and big data in supply chain management: Current state and future potential. *Journal of Business Logistics, 36*(1), 120-132.

Taiwo, O. A. (2010). Types of Machine Learning Algorithms. Dans Y. Zhang (Éd.), *New Advances in Machine Learning* (pp. 3 – 31). Portsmouth, United Kingdom: InTech, University of Portsmouth.

Tsourgiannis, L. N., & Valsamidis, S. I. (2021). Clustering and Profiling Consumer Buying Behavior: The Greek Case During Crisis Period. *International Journal of Knowledge-Based Organizations*, 11(1), 50-66.

Ullah, I., Raza, B., Malik, A., Imran, M., Islam, S., & Kim, S. (2019). A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. *IEEE Access, 7*, 60134-60149.

Visani, F., & Boccali, F. (2020). Purchasing price assessment of leverage items: A Data Envelopment Analysis approach. *International Journal of Production Economics*, 223, 107521.

Von Kirby, P., Gerardo, B. D., & Medina, R. P. (2017). Implementing enhanced AdaBoost algorithm for sales classification and prediction . *International Journal of Trade, Economics and Finance*, 8(6), 270-273.

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, 165, 234-246.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1), 67-82.

Wuest, T., Irgens, C., & Thoben, K. (2014). An approach to monitoring quality in manufacturing using supervised machine learning on product state data. *Journal of Intelligent Manufacturing, 25*(5), 1167-1180.

Xu, Y., Balakrishnan, S., Singh, A., & Dubrawski, A. (2020). Regression with comparisons: Escaping the curse of dimensionality with ordinal information. *The Journal of Machine Learning Research, 21*(1), 6480-6533.

Xu, Y., Jones, G. J., Li, J., Wang, B., & Sun, C. (2007). A study on mutual information-based feature selection for text categorization. *Journal of Computational Information Systems*, 3(3), 1007-1012.

Zhang, Y., Minchin Jr, R., & Agdas, D. (2017). Forecasting completed cost of highway construction projects using LASSO regularized regression. *Journal of Construction Engineering and Management-ASCE, 143*(10).

Zhong, R. Y., Huang, S., Lan, Q. Y., Dai, X., Chen, & Zhang, T. (2015). A big data approach for logistics trajectory discovery from RFID-enabled production data. *International Journal of Production Economics, 165*, 260-272.