AI-Based Career Transition Recommendation System Using Controlled Educational Data Augmentation

Gerlix ADANKON[1], Pelagie HOUNGUE[2], Melckior DEGBOE[3], Corelle GOGAN[4]
L@RIAD Laboratory of Institute of Mathematics and Physical Sciences (IMSP), University of Abomey-Calavi, Dangbo, Benin[1]
LITA Laboratory of Higher Polytechnic School (ESP), University Cheikh Anta Diop (UCAD), Dakar, Senegal[3, 4]

***Abstract*—Career transition into digital professions is a strategic lever for addressing youth unemployment in Sub-Saharan Africa. However, existing training programs lack objective career guidance tools. This study presents two complementary contributions, evaluated using data collected from 131 professionals who underwent a career transition process into a digital profession. Learn-Orient is a career path recommendation system combining a profile recognition filter based on Gower distance normalized by the IQR and a binary logistic regression model, producing probabilistic estimates with a graded confidence level (High, Moderate, Low). EDU-CDA is a data augmentation method suited for small educational datasets with mixed variables and imbalanced classes, combining conditional SMOTE for continuous variables and conditional Bernoulli sampling for binary variables. Validated by a three-way protocol (distributional overlap 82.9–90.8%, TVD < 0.061, TSTR Δ = 0.025), EDU-CDA transforms the logistic regression model, which is typically the most vulnerable to small sample sizes (AUC = 0.810 in real cross-validation)—into the most stable and high-performing one (AUC = 0.995 in cross-validation, F1 = 0.900 on a real test set of 20 observations). Among the 16 pre-training predictors selected, institutional funding (OR = 32.40) proves to be the most discriminating factor for the Developer profile, while the creativity test score (OR = 0.107) strongly characterizes the Designer profile. The system incorporates a filter for detecting atypical profiles, ensuring responsible use in a decision-making context with significant human stakes.**

Keywords*—Machine learning; data augmentation; predictive model; artificial intelligence in education; professional retraining; career recommendation

I. INTRODUCTION

Choosing the right educational track is one of the most critical decisions in a professional career. Yet, this choice is still too often based on subjective factors such as personal intuition, family influences, and opportunities available at a given time, rather than on tangible data and objective profile assessment tools. The consequences of inappropriate academic guidance are well-documented and severe. This is what [1] reveal in their study reveals that 51.2% of students at the Autonomous University of Santo Domingo report having chosen the wrong program, and 85.6% have considered dropping out. Wydra-Somaggio [2], meanwhile, they found that 70% of learners who prematurely terminated an apprenticeship contract in Germany pursue a career change. These findings all point to the fact that the mismatch between a candidate's profile and the chosen program is a major determinant of dropout rates and the waste of resources invested in education.

This issue is even more profound and significant in sub-Saharan Africa. With a youth unemployment rate estimated at 12.7% [3] and a persistent mismatch between available skills and the actual needs of the labor market [4], the region faces a structural challenge in terms of labor market integration. In this context, digital professions represent a strategic opportunity. In fact, 230 million digital jobs are expected by 2030 [5]. Professional retraining programs, such as short, intensive training courses designed for candidates from diverse backgrounds, serve as a direct pathway to these jobs. However, their effectiveness faces a persistent obstacle: the lack of objective tools capable of guiding each candidate toward the track best suited to their profile to maximize their chances of success [6]. Moreover, these challenges are amplified by local specificities and constraints, such as limited educational infrastructure [7], the complex situation of professionals undergoing retraining who are forced to keep their jobs while training without any support, and a digital labor market that is still taking shape.

Machine learning offers a promising solution to this challenge [8]. The systematic review by Trujillo et al. [9], covering 38 studies selected from 1,296 articles, demonstrates that machine learning (ML)-based career recommendation systems are rapidly developing in higher education. However, these systems rely almost exclusively on standardized academic data from large datasets in North American or European contexts, making them ill-suited to the heterogeneous context of career transition in Sub-Saharan Africa, where profiles are multidimensional, skills are self-reported, and available data is both low in volume and imbalanced across categories [9]. The concept of low volume is particularly relevant given that, in the African regions, data collection frameworks have not been established to regularly provide reliable information from projects, programs, and training pathways focused on professional retraining. Furthermore, standard data augmentation methods such as SMOTE [10] prove to be ineffective on predominantly binary variables with logical constraints [11]. Few studies propose an approach capable of simultaneously modeling this richness of profiles and addressing these specific methodological constraints. This study is grounded in this dual gap, both practical and scientific.

This article proposes two original and complementary contributions, evaluated on the complete dataset collected from 131 candidates, including 99 observations (or obs) used, and on augmented data ranging from 200 to 20,000 observations.

(C1) Learn-Orient is a major-track recommendation system combining a profile recognition filter based on the Gower

distance normalized by the IQR (Interquartile Range) and an interpretable binary logistic regression model, producing probabilistic estimates with a graded confidence level. It currently enables guidance toward the two professions of graphic designer and application developer. (C2) EDU-CDA (Educational Controlled Data Augmentation) is a controlled augmentation method tailored to small educational datasets with mixed variables and imbalanced classes, combining conditional SMOTE for continuous variables and conditional Bernoulli sampling for binary variables, validated by a tripartite protocol. Four research questions structure this approach:

RQ1 — Which pre-training characteristics (socio-demographic, motivational, and technical) enable the ability to distinguish candidates oriented toward the Graphic Design track from those oriented toward the Application Developer track, and with what predictive performance?

RQ2 — To what extent does a controlled augmentation method adapted to mixed variables and small sample sizes improve the predictive performance of classification models compared to models trained exclusively on real data?

RQ3 — Do the synthetic data generated by EDU-CDA exhibit sufficient statistical fidelity to the real data, as measured by a three-part validation protocol that combines distributional fidelity, proportional similarity, and predictive utility?

RQ4 — Is the Gower distance normalized by the IQR a reliable indicator of a candidate profile's proximity to the known data space, enabling the identification of atypical profiles for which automatic prediction is unreliable?

To answer these questions, the remainder of the article is organized as follows: the next section presents the literature review and state of the art; the methodology is then described; the results are presented; the following section discusses the implications, while the final section concludes and offers recommendations.

II. Related Work

A. African Context and Challenges of Digital Reskilling

Professional retraining for digital careers is emerging in Sub-Saharan Africa as a structural response to the mismatch between the supply of available skills and the actual needs of the labor market [4]. In this context, short-term training programs for candidates from diverse backgrounds are proliferating, but their effectiveness remains limited by the lack of objective career guidance tools. Macharia [6] emphasizes that retraining candidates frequently lack precise information about available career paths and their actual requirements. To be effective in this context, smart educational technologies must be designed with African challenges and the specific context of professional retraining in this region in mind, including challenges such as the coexistence of work and training, which reduces learners' cognitive capacity to devote to their training, and the heterogeneity of incoming profiles. It is precisely this heterogeneity that renders approaches developed in contexts with abundant, standardized academic data inadequate.

B. Machine Learning for Career Guidance

The application of machine learning to academic and career guidance has been the subject of a growing body of scientific research. Trujillo, Pozo, and Suntaxi [9] provide the most recent and comprehensive synthesis, with a systematic review of 38 studies selected from 1,296 articles. Their results show that random forests (26% of studies), SVMs (21%), and neural networks (16%) dominate the field, while logistic regression (13%) is preferred in contexts where the interpretability of recommendations is paramount. Academic data constitute the most frequent input variable (40% of studies), followed by personal interests (25%) and demographic data (15%). F1-score, precision, and recall are identified as the most used evaluation metrics, justifying their consideration in this research methodological choices. The authors, however, identify two persistent structural limitations: the unavailability of open and reproducible datasets, which hinders comparisons between studies, and the near absence of longitudinal evaluations to measure the actual impact of recommendations on career outcomes. These gaps are particularly pronounced in African contexts, where education data remains fragmented and rarely shared publicly.

In the specific field of IT tracks, Al-Dossari et al. [12] developed CareerRec using data from 2,255 IT employees in Saudi Arabia, comparing five algorithms to predict membership in three professional tracks. Despite a sample twenty times larger than the sample currently used for this study, the maximum accuracy achieved, 70.47% with XGBoost, highlights the inherent difficulty of predicting IT career paths. It should also be noted that this study does not specifically address career transition. Two contributions of this study are directly applicable. These are the binarization of programming language skills into independent binary columns, identical to the multi-label encoding used for this study, and the confirmation that class imbalance significantly degrades predictive performance and requires dedicated handling.

In the Beninese context, Houngue, Hountondji, and Dagba [13] developed a post-BEPC guidance system achieving 99% accuracy on 325 students using random forests. This result, obtained using standardized and objectively measurable academic scores, demonstrates the technical feasibility of supervised approaches in the Beninese educational context. However, it is not transferable to our configuration, where the dataset of 99 observations draws on self-reported skills, motivations, and economic context factors, thus operating within a framework that is both more complex and less structured, with a smaller sample size. Wu et al. [14] confirm, moreover, that variable selection and class balancing constitute the two most critical methodological challenges once one moves beyond standard academic settings.

C. Interpretability and Choice of Production Model

The choice of classification model is not merely a matter of maximizing a performance metric. Rudin [15] presents a decisive normative argument stating that in fields with high human stakes, including career counseling, interpretable models are preferable to black-box models when they offer comparable performance, as they allow for the detection of bias, the justification of decisions, and the maintenance of user trust.

Logistic regression meets this requirement, and its coefficients translate directly into odds ratios, providing a transparent justification for each recommendation generated [16]. Song et al. [17] extend this argument by showing that incorporating motivational and behavioral variables, beyond academic characteristics alone, improves the accuracy of career trajectory predictions, which forms the basis of the selected predictive variable structure for this study. Smirani et al. [18], Lee et al. [19], and Arévalo-Cordovilla and Peña [20] further confirm that ensemble models suffer from overfitting on small samples and exhibit high sensitivity to hyperparameters, two undesirable properties in this context of working with 99 actual observations.

D. Class Imbalance and Data Augmentation

The problem of learning on imbalanced data is well-documented. He and Garcia [11] identify two families of solutions: data-level methods, including synthetic resampling, and algorithmic-level methods such as class weighting. They emphasize that class imbalance interacts with the small sample size to jointly degrade the models' sensitivity to minority classes, a situation encountered in the dataset, which has 38 Developers versus 61 Designers.

According to the literature, SMOTE [10] is the most widely adopted resampling method. It creates synthetic observations via linear interpolation between a real observation and one of its k nearest neighbors in the feature space, producing greater diversity than simple repetition-based oversampling. Anggrawan et al. [21] and Cu et al. [22] confirm its effectiveness in imbalanced educational contexts, with sensitivity gains of up to 25%. However, SMOTE has two critical limitations for modeling use. Indeed, when applied to binary variables, it generates decimal values with no semantic meaning (e.g., lang_java = 0.5). Furthermore, since it is not class-conditional, it risks producing hybrid profiles across tracks on an unbalanced dataset. These specific limitations further motivate the development of EDU-CDA.

E. Validating the Quality of Synthetic Data

Evaluating the quality of synthetic data requires additional criteria beyond descriptive statistics alone. Esteban et al. [23] formalized the TSTR (Train on Synthetic, Test on Real) protocol. This involves a model trained exclusively on synthetic data and evaluated on real data. The difference from the model trained on real data (TRTR) measures the predictive utility of the synthetic data [24]. Jordon et al. [25] confirm that TSTR is currently the gold standard in this field, as it evaluates the preservation of relational structures between variables, rather than simply marginal distributions. Qian et al. [26] additionally propose the Total Variation Distance (TVD) within the Synthcity framework. They suggest calculating it variable by variable as the absolute difference between the proportions observed in the real and synthetic data. It constitutes a non-parametric metric directly applicable to binary variables. The threshold of 0.05 used as the limit for a negligible difference is consistent with the conventional threshold for statistical significance. Specifically in the educational context, a recent benchmark [27] comparing SMOTE and several deep generative models on 10,000 students shows that resampling methods achieve a TSTR of 0.997, confirming their predictive utility on tabular data. A TSTR score of 0.924, obtained on a dataset twenty times smaller and with a higher proportion of constrained binary variables, is consistent with this benchmark.

F. Handling Missing Values and Variable Selection

Missing value handling must precede any modeling of field data. Batista and Monard [28] have empirically demonstrated that KNN imputation produces statistically superior results compared to methods based on the mean or mode, as it leverages local relationships between variables to estimate missing values. Li et al. confirm that including the target variable in the search for neighbors improves the quality of imputation by leveraging available class information, a practice adopted in this pipeline. Emmanuel et al. [29] justify the exclusion of variables with more than 30% missing values, a threshold beyond which imputation introduces more noise than useful information.

Variable selection is a critical issue when the number of variables far exceeds the number of observations. Guyon and Elisseeff [30] have shown that combining several complementary methods reduces the risk of eliminating important variables due to the lack of power of a single method, justifying the three-method voting system with a unanimity requirement used in a context of $n = 99$ observations of the dataset [31]. Multicollinearity checking completes this step by considering the work of Dormann et al. [32] and Hair et al. They recommend a threshold of $|r| > 0.70$ as an exclusion criterion, above which two variables share more than 49% of their variance, making the regression coefficients unstable and their interpretation unreliable.

G. Evaluation Metrics and Classification Threshold

In the presence of imbalanced classes, the choice of metrics is a methodological decision. Japkowicz and Shah [33] demonstrate that a classifier that systematically predicts the majority class would achieve a high accuracy rate in the dataset without any discriminative value, rendering this metric inappropriate. The F1-score, the harmonic mean of precision and recall, is the primary recommended metric because it penalizes both types of errors equally. The AUC [34] provides an evaluation of overall discriminative ability independent of the threshold, which is particularly stable in cross-validation on small folds.

Determining the optimal classification threshold is distinct from model selection. The Youden index [35] simultaneously maximizes both detection rates under the assumption of symmetric costs between false positives and false negatives. In the considered career guidance context, this assumption is methodologically justified, since mistakenly directing a Developer toward Design and mistakenly directing a Designer toward Development represent equally costly failures for the candidate.

H. Similarity Measure for Mixed Variables and Person-Environment Fit Theory

Measuring similarity between profiles with continuous and binary variables requires an appropriate metric. Gower [36] proposed a general coefficient calculated as the average of individual similarities normalized according to the type of each variable, natively handling missing data and mixed types. D'Orazio [37] identifies a limitation of the standard version, noting that continuous variables with a wide interquartile range

tend to dominate the overall calculation. He proposes normalizing by the IQR rather than by the total range, thereby mitigating the impact of outliers and balancing the contribution of both types of variables. This distance is used in a Learn-Orient as a profile recognition filter before any prediction.

This filter is theoretically grounded in the concept of person–environment fit, formalized by Holland [38], which posits that professional satisfaction and success depend on the congruence between an individual's characteristics and the demands of their environment. This framework underpins the hypothesis that variables such as creativity scores or prior technical skills constitute stable indicators of natural fit for a particular field. Jaoui and Hilmi [39], in the French-speaking context of career transition, confirm that these fit factors remain robust determinants of career change decisions, regardless of formally acquired skills.

I. Positioning of the Present Study

The literature review presented in this chapter highlights four complementary gaps.

To begin with, the absence of a system for predicting and recommending career paths in the context of adult learning and career transition, particularly toward digital professions.

Second, existing work on career prediction using ML operates almost exclusively in contexts with large datasets containing standardized academic variables. No study offers a methodological solution for the simultaneous occurrence of a small sample size, predominantly binary variables with logical constraints, and imbalanced classes, a configuration that is characteristic of career transition data, particularly in an African context where data scarcity is also a common phenomenon.

Furthermore, while SMOTE has been validated for data augmentation on continuous educational data, its application to constrained binary variables remains unaddressed in the literature. The EDU-CDA method proposed in this study directly fills this gap by combining conditional SMOTE for continuous variables and conditional Bernoulli sampling for binary variables.

Finally, there is a lack of a component for detecting atypical profiles in existing guidance systems. No existing guidance system distinguishes between candidates whose profiles are covered by the model and those for whom the prediction is unreliable. The integration of the Gower distance normalized by the IQR as a pre-prediction filter constitutes an operational contribution that addresses this limitation.

These four gaps collectively define and establish the scope of this study, whose methodology is presented in the following chapter.

III. METHODOLOGY

A. Sampling and Data Collection Strategy

The target population of this study, individuals seeking career transition into digital professions, is not tracked by any official registry in Benin. No institution is formally dedicated to digital career transition in the strict sense. The phenomenon manifests itself in a diffuse manner across various training institutions (private digital schools or universities, vocational training centers, and integration programs) where some learners change careers or fields of study without this process being institutionally identified as such. The sampling universe is therefore, by nature, not delimited a priori, rendering any probabilistic approach inapplicable in the absence of an exhaustive sampling frame.

Considering this structural constraint, a non-probabilistic network sampling strategy, commonly referred to as snowball sampling, was adopted [40]. This method is recognized as appropriate when the target population is difficult to reach, unregistered, or distributed across multiple structures without administrative centralization [41], [42]. It involves identifying an initial group of participants who meet the inclusion criteria, then asking them to identify other potential candidates within their networks.

In the case of this study, the process began with direct contact with several institutions offering training in digital professions in Cotonou and surrounding cities, identified as likely to accept learners seeking career transition. A 47-item Google Forms questionnaire was distributed to these institutions and then shared by the initial respondents with their peers.

Two methodological precautions were taken to limit the biases inherent in this approach. At the outset, the deliberate diversification of entry points into the network. Several distinct institutions were officially contacted simultaneously to reduce the risk of overrepresentation of a particular subgroup linked to a single institution or a single social network. In addition, the anonymous and self-administered nature of the form limits social desirability bias in responses to questions about motivations and self-reported skills. Completing the form was voluntary, with no pressure from the administration. Respondents were not under any influence.

B. Data Collected and Sample Characteristics

This approach yielded 131 responses, of which 99 were retained after applying eligibility criteria that excluded programs with too few students for defensible statistical modeling (Audiovisual Specialist: 11 obs; Digital Project Manager: 8 obs; Data Analyst: 1 obs) and programs with ambiguous or incomplete titles. The primary inclusion criterion is the presence of an identifiable career break, where the candidate must have worked in or prepared for a career in a field distinct from digital professions before enrolling in their current program. This criterion operationalizes the concept of career transition regardless of the institution attended. The final dataset includes 61 candidates (61.6%) in the Graphic Design track and 38 (38.4%) in the Application Developer track, representing a 1.6:1 imbalance ratio. Table I presents the main characteristics of the sample.

C. Methodological Pipeline

The methodology is organized into seven sequential, interdependent steps, summarized in Table II. The dataset is divided into a training set (80%, 79 obs) and a test set (20%, 20 obs) via random sampling with a fixed seed. The test set is strictly isolated from the training data. It is not used in either variable selection or the construction of the EDU-CDA synthetic data, ensuring an unbiased evaluation.

TABLE I. SAMPLE CHARACTERISTICS BY TRACK (N = 99)

Characteristic	Graphic Designer (n=61)	Application Developer (n=38)
Average Age (years)	23.9 (σ=4.6)	24.6 (σ=5.4)
Male	67.2%	78.9%
From Benin	95.1%	97.4%
Family funding	70.5%	57.9%
Institutional funding	6.6%	18.4%
Personal funding	23.0%	23.7%
High school diploma or higher	34.4%	47.4%
Previous experience in the digital field	29.5%	50.0%
Availability > 72 hours/week	14.8%	26.3%
Average work experience	28.4 months	38.2 months

TABLE II. EDU-CDA METHODOLOGICAL PIPELINE

Step	Title	Input	Output
P1	Preprocessing & normalization	131 observations, 49 raw variables	99 obs, clean variables
P2	Variable coding	Textual/categorical variables	76 numerical variables
P3	Feature engineering	76 variables	80 variables (+ 4 derived variables)
P4	Variable selection (voting among 3 methods)	79 candidate variables	16 final variables
P5	Baseline modeling (real data)	79 real training observations	Baseline performance
P6	EDU-CDA increase	79 real training observations	1,600 balanced observations
P7	Final modeling + Evaluation	1,600 augmented observations	Evaluation of 20 real test observations

D. Preprocessing, Encoding, and Feature Engineering

Preprocessing begins with converting ambiguous textual responses (none, nothing, n/a, none, nr, nc, etc.) into missing values. The variable moy_bepc, which has an N/A rate of 35.4%, is excluded in accordance with the 30% threshold recommended by Emmanuel et al. [29]. The four remaining numerical variables (moy_bac, test_creativity, note_dipl, experience_month) are imputed using KNN (k = 5) with the target variable included in the distance calculation, which leverages the available class information [28].

The encoding transforms the textual variables using four complementary approaches: 1) simple binary encoding (gender, internet access, Beninese origin); 2) one-hot encoding with a priori categorical grouping for five variables, city (4 geographic zones), education level (Bac_Less / Bac_More), funding (4 types), field (3 categories), sector (3 categories); 3) multi-label encoding using regular expressions for multiple-choice variables (programming languages, design tools, office and cross-functional skills); (iv) lexical field detection for free-text responses (motivation, objective).

Four derived variables are constructed to synthesize the information scattered across individual binary columns with small sample sizes, improving the statistical power of the selection tests applied in the next step [30]:

$$\text{profile\_dev} = \Sigma\ \text{programming\_languages} + \text{html\_level} \quad (1)$$

$$\text{designer\_profile} = \Sigma\ \text{design\_tools} + \text{photoshop\_level} \quad (2)$$

$$\text{nb\_trans} = \Sigma\ \text{cross-functional\_skills} \quad (3)$$

$$\text{nb\_office} = \Sigma\ \text{office\_tools} \quad (4)$$

E. Selection of Variables by Unanimous Vote Using Three Methods

The selection is based on 79 variables available after feature engineering, using a voting system with three complementary methods [30]. A variable is retained only if it simultaneously satisfies all three criteria (score 3/3), which reduces the risk of spurious correlation in a sample of 99 observations [31]:

One-vs-all correlation: $r_j = \left|cor\left(X_j, Y_k\right)\right| > 0.10$, where Y_k is the binary indicator for sector k.

Random Forest Importance (MDA): 500 trees, average decrease in accuracy when permuting values > 0.

Statistical tests: Kruskal-Wallis (continuous variables); Fisher's exact test (binary variables, small sample sizes); $p < 0.05$.

Checking for multicollinearity at a threshold of $|r| > 0.70$ [33] led to the exclusion of 8 redundant variables as presented in Table III. The most informative variables within each set of redundant variables were retained.

TABLE III. EXCLUDED REDUNDANT PAIRS (|R| > 0.70)

Excluded variable	Retained variable	Observed |r|
high_school_level	high_school_level	1.000
html_level	dev_profile	0.928
Photoshop level	designer_profile	0.925
trans_proact	nb_trans	0.780
lang_html	dev_profile	0.775
illustrator_tool	designer_profile	0.765
no_trans	nb_trans	0.762
no_lang	dev_profile	0.752

The 16 final variables selected are presented in Table IV.

TABLE IV. 16 FINAL VARIABLES SELECTED (STEP P4)

Variable	Type	Category / Description
Age	Continuous	Socio-demographic
work_style	Continuous	Motivational (preference for solo vs. group)
logic_test	Continuous	Skill (logic test score)
creativity_test	Continue	Skill (creativity test score)
dev_profile	Continue	Derived — Σ languages + HTML level
designer_profile	Continue	Derived — Σ design tools + photoshop_level
nb_trans	Continue	Derivative — Σ cross-disciplinary skills

post-high-school level	Binary	Level of education ≥ High school diploma
end_of_family	Binary	Family funding
end_institutional	Binary	Institutional funding (scholarship)
time-intensive	Binary	Availability > 72 hours/week
Digital_Tech	Binary	Previous field: digital/tech
dom_other	Binary	Previous field: other
lang_java	Binary	Prior knowledge of Java
lang_js	Binary	Prior knowledge of JavaScript
trans_redac	Binary	Declared writing skills

F. EDU-CDA Method – C2 Contribution

The complete augmentation process is detailed in Fig. 1.

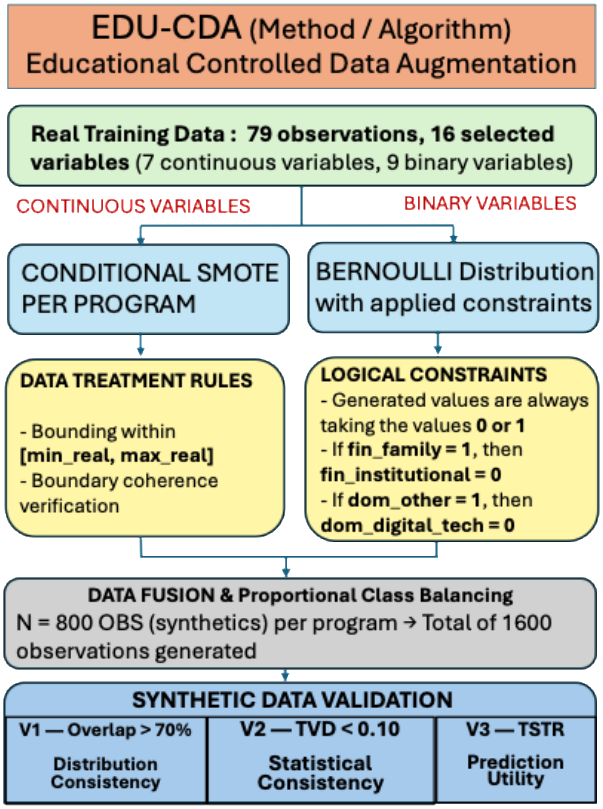

Fig. 1. EDU-CDA process.

EDU-CDA (Educational Controlled Data Augmentation) is one of this study's main contributions, developed to simultaneously address the dataset's three constraints: low volume (79 training observations), predominantly binary variables with logical constraints (9 out of 16), and imbalanced classes (1.6:1 ratio). It combines two generation mechanisms applied separately for each track k ∈ {Designer, Developer}, avoiding the generation of hybrid profiles at the boundary between the two classes.

1) Class-conditional SMOTE (continuous variables): For each track k, synthetic observations are generated by linear interpolation between a real observation x_i and one of its K nearest neighbors x_{nn} within the same class [10]:

$$\mathbf{x_{synth}} = x_i + \lambda \cdot (x_{nn} - \mathbf{x_i}),\ \lambda \sim U(0, 1) \quad (5)$$

The parameter K is adjusted: $K = \min(3, n_k - 1)$, avoiding outliers in classes with small sample sizes [11]. Two post-processing steps are systematically applied: 1) clamping values within the observed bounds $[x_{min}, x_{max}]$; 2) rounding to the nearest integer for count variables (profil_dev, profil_designer, nb_trans, test_logic, test_creativity, work_style).

2) Conditional Bernoulli per class (binary variables): For each binary variable x_j and each track k, the probability of occurrence is estimated from the actual training data:

$$\hat{p}_{jk} = \frac{1}{n_k} \cdot \sum_{i:y_i=k} x_{ij} \quad (6)$$

Each synthetic value is generated by a Bernoulli draw:

$$x_{ij}^{synth} \sim \text{Bernoulli}(\hat{p}_{jk}) \quad (7)$$

This approach guarantees values strictly ∈ {0, 1}, preserving the binary nature of the variables. It is preferred over SMOTE for binary variables whose interpolation would produce decimal values that are semantically unacceptable in the context of self-reported skills.

3) Consistency rules and augmentation strategy: Two business consistency rules correct logically impossible combinations: 1) if fin_family = 1, then fin_institutional = 0; 2) if dom_other = 1, then dom_digital_tech = 0.

Two augmentation strategies were used to determine the optimal strategy. The balanced strategy generates N synthetic observations per class regardless of the actual volume, resulting in approximately equal sample sizes across fields. The proportional strategy, on the other hand, generates $(N - n_k)$ synthetic observations per track, where n_k denotes the number of actual observations in class k, thereby ensuring exactly N observations per class in the final augmented dataset.

G. Validation of Synthetic Data

The quality of the synthetic data is evaluated according to three independent criteria, covering the distributional fidelity of continuous variables, the statistical fidelity of binary variables, and the predictive utility of the synthetic dataset.

1) Distributional fidelity (continuous variables): For each continuous variable, the overlap between the distributions of the real and synthetic data is calculated by intersecting the normalized histograms:

$$Overlap_j = \frac{\Sigma_b \left(h_j^{real}(b), h_j^{synth}(b)\right)}{\Sigma_b\, h_j^{real}(b)} \quad (8)$$

where, hj (b) denotes the density of variable j in bin b. An overlap greater than 70% is used as the acceptability criterion.

2) Statistical fidelity (binary variables): For each binary variable, the Total Variation Distance (TVD) between the real and synthetic proportions is calculated, Qian et al. [26]:

$$TVD_j = \left|\widehat{p_j^{réel}} - \widehat{p_j^{synth}}\right| \quad (9)$$

A TVD of less than 0.05 is considered negligible, by analogy with the conventional significance threshold α=0.05. A TVD of less than 0.10 is considered acceptable.

3) Predictive Utility (TSTR): An LDA model is trained exclusively on synthetic data and evaluated on real training data (TSTR), in accordance with the protocol by Esteban et al. [23]. Its accuracy performance is compared to that of an LDA model trained on real data only (TRTR). The choice of LDA as the reference classifier for this protocol is motivated by its stability across small samples. A difference of less than 0.15 is considered acceptable [43]:

$$\Delta_{TSTR} = |AccTRTR - AccTSTR| < 0{,}15 \quad (10)$$

H. Learn-Orient System - Contribution C1

The Learn-Orient system combines two sequential steps to produce a major recommendation with a graded confidence level. Fig. 2 illustrates the complete workflow.

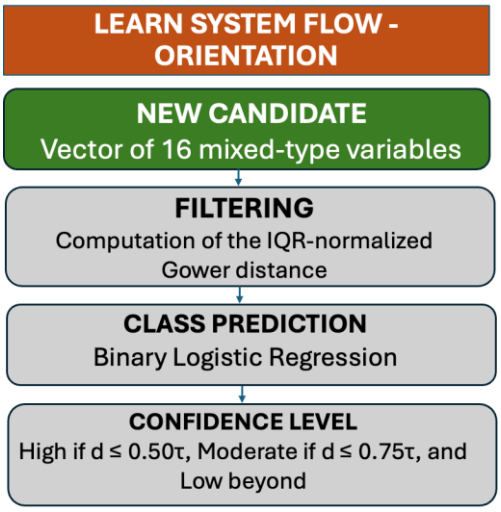

Fig. 2. Workflow of the learn-orient guidance system.

1) Profile recognition filter: The Gower distance normalized by the IQR [37] is calculated between the new candidate's profile and each of the 1,600 reference observations. For continuous variables, partial similarity is normalized by the IQR (rather than the range) to mitigate the impact of outliers and balance the contribution of both types of variables. For binary variables, the asymmetric Jaccard coefficient is used so that a double absence does not contribute to the similarity. The average distance to the k = 5 nearest neighbors is calculated as follows:

$$\underline{d}_G ower(x_n ew) = (1/5) \cdot \Sigma_{j=1}^{5} d_G ower\left(x_n ew, x_{(j)}\right) \quad (11)$$

A threshold τ is defined as the 95th percentile of the distribution of mutual distances between reference observations. If $\bar{d} > \tau$, the profile is declared unrecognized, and the system recommends considering other digital career paths besides designer and developer, thereby signaling its own limitations for atypical profiles (RQ4).

2) Binary logistic regression and confidence levels: If the profile is recognized, binary logistic regression estimates the membership probabilities:

$$P(x) = 1/\left(1 + exp\left(-(\beta_0 + \beta^T x)\right)\right) \quad (12)$$

$$P(Designer \mid x) = 1 - P(Developer \mid x) \quad (13)$$

The β coefficients are estimated using the augmented dataset (N = 1,600) and interpreted as odds ratios $OR_j = \exp(\beta_j)$ [34]. An odds ratio greater than 1 indicates a positive association with the Application Developer track; an odds ratio less than 1 indicates a positive association with the Graphic Designer track. At the same time, confidence intervals are evaluated to identify the variables significant for predicting the targeted tracks. If the interval $[CI_{lower}, CI_{upper}]$ contains 1, the variable is considered insignificant.

The optimal threshold is determined by the Youden index [35]: J = Sensitivity + Specificity − 1. Three confidence levels are assigned based on $\bar{d} / \tau$:

- High (≤ 0.50), reliable automatic prediction;
- Moderate (≤ 0.75), recommendation for further evaluation;
- Low (> 0.75 but ≤ 1.00), maintenance strongly recommended [10].

I. Algorithms

Seven binary classification algorithms are compared in two phases: a baseline phase on real-world data (P5) and a post-augmentation phase on EDU-CDA data (P7). The evaluated algorithms are binary logistic regression (LR), logistic regression with class weighting (LR + Weighting), radial basis support vector machine (SVM), random forest (RF), linear discriminant analysis (LDA), naive Bayes classifier, and decision tree (CART).

The dataset is divided into a training set (80%, or 79 observations) and a test set (20%, or 20 observations) via simple random sampling with a fixed seed (set.seed(42)). The test set is used only once, for the final evaluation, and is not involved at any stage in the construction of synthetic data or in variable selection.

J. Evaluation Metrics

1) Supplementary metrics: Five complementary metrics are calculated for each model. Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively, with the positive class being the Application Developer track.

Accuracy: Overall proportion of correct predictions, provided for informational purposes only. In the presence of imbalanced classes, a classifier that systematically predicts the majority class can achieve high accuracy without discriminative value [33].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (14)$$

Sensitivity and Specificity: Sensitivity measures the proportion of correctly identified Application Developers; specificity measures the proportion of correctly identified Designers:

$$Sensitivité = \frac{TP}{TP+FN} \qquad Spécificité = \frac{TN}{TN+FP} \tag{15}$$

F1-score: Primary model selection metric, defined as the harmonic mean of precision and recall, suitable for imbalanced problems [33]:

$$F1 = 2 \cdot \frac{Precision \times Sensitivity}{Precision + Sensitivity} \tag{16}$$

AUC: Area under the ROC curve, measuring overall discriminatory power independently of the threshold [34]. Primary metric in cross-validation due to its stability on small folds. Train-Test Gap: An indicator of overfitting [44]:

$$\Delta_{overfitting} = F1_{train} - F1_{test} \tag{17}$$

A significant discrepancy between training and test performance indicates that the model has memorized the data rather than learned generalizable patterns.

2) Cross-validation: Three complementary 5-fold cross-validations are conducted using the caret package.

CV1 — On real data. A 5-fold cross-validation is performed on the 79 real training observations to estimate the performance of each algorithm more robustly than a single split, whose random composition on a small sample can produce unstable estimates. AUC is chosen as the evaluation metric due to its stability on small folds.

CV2 — On augmented data. A 5-fold cross-validation is performed on the 1,600 augmented observations to assess the stability of the final model's performance on the synthetic data space. Both cross-validations are implemented using the `trainControl`.

CV3 — On real data after augmentation. A third cross-validation complements CV1 and CV2 to quantify the reliability of the performance estimates and to test the significance of the differences between models. CV3 is a stratified, repeated cross-validation (5 folds, 10 repetitions, i.e., 50 evaluations per model). Unlike CV1, which trains on the real data alone, and CV2, which trains and tests on the augmented data, EDU-CDA augmentation is applied only to the training partition of each fold, the test fold remaining composed exclusively of real observations. The distribution of the 50 F1 scores obtained per model is used to estimate a 95% confidence interval. Finally, the differences between models are tested using the exact McNemar test applied to the out-of-fold predictions: for each pair of models, the test examines only the observations on which the two models disagree and checks whether these disagreements lean significantly in favour of one or the other.

K. Tools and Technical Environment

The analyses were performed using R (version 4.x) and Python 3.x via Google Colaboratory. In R, the main packages used were: caret (modeling and cross-validation), VIM (KNN imputation), ranger (random forests), glm (logistic regression), StatMatch (Gower's distance), and pROC (ROC curves and AUC). In Python, the scikit-learn, imbalanced-learn, pandas, and numpy libraries were used for preprocessing and algorithm comparison. Reproducibility is ensured by setting the random seeds (set.seed(42) in R, random_state=42 in Python).

IV. Results

A. Variable Selection (RQ1)

Out of 79 candidate variables, the unanimous vote across three methods selected 24 variables. As shown in Fig. 3, after checking for multicollinearity, 16 final variables were selected (Table IV). The correlation heatmap confirms the overall independence of the 16 final variables. The strong correlations observed are expected and consistent: fin_famille and fin_institutionnel show a negative correlation (mutually exclusive sources). test_logique and test_creativite show a moderate positive correlation. profil_dev and profil_designer do not have a strong correlation between them, confirming their ability to distinguish the two tracks independently. The absence of strong correlations among the other variables validates the quality of the P4 selection.

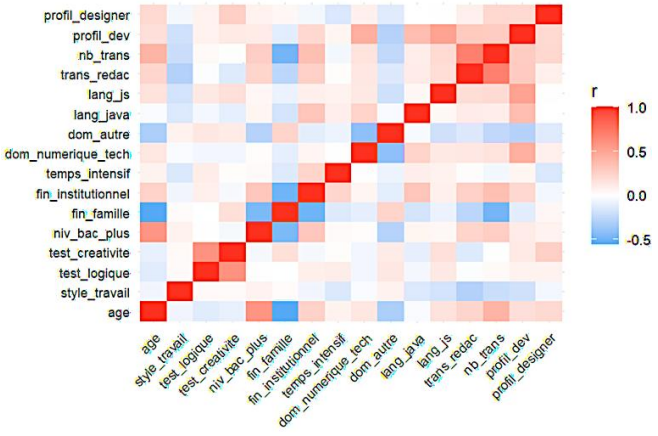

Fig. 3. Heatmap of 16 selected variables.

Profil_dev has the highest correlation with the target (r = 0.427) and the most pronounced difference in means (Designer: 1.13 vs. Developer: 3.13). profil_designer shows the inverse relationship (r = −0.291). These derived variables effectively synthesize the information scattered across individual binary columns with small sample sizes.

B. Validation of EDU-CDA Synthetic Data (RQ3)

The three criteria for tripartite validation are fully satisfied, providing a positive answer to RQ3 as shown in Table V.

TABLE V. Tripartite Validation of EDU-CDA Synthetic Data

Criterion	Indicator	Threshold	Result	RQ3
Distribution accuracy	Histogram overlap (7 continuous variables)	> 70%	82.9–90.8%	Yes
Statistical accuracy	TVD proportions (9 binary variables)	< 0.10	Max. = 0.061	Yes
Predictive utility	TSTR: ΔAccuracy LDA (synthetic vs. real)	Δ < 0.15	Δ = 0.025	Yes

C. Comparative Performance of the Models (RQ1, RQ2)

Table VI and Fig. 4 present the performance of the seven algorithms on the real test set. In the baseline phase (P5), LDA is the model with the best F1 score (F1 = 0.889, AUC = 0.940, ΔTrain = 0.060). Binary LR and RF achieve a perfect training score (1.000), indicating severe overfitting due to near-linear separability in a 16-variable space across 79 observations [38]. After EDU-CDA augmentation (P7), Binary LR is the only model to achieve F1 = 0.900 while maintaining the lowest ΔTrain among the competitive models (0.080), positively addressing RQ2. RF retains persistent overfitting (ΔTrain = 0.222). The Binary LR is selected as the production model.

TABLE VI. PERFORMANCE ON REAL TEST SET (20 OBS) ON P5 AND P7

Model	Acc.	F1	Sens.	Spec.	AUC	Ph	ΔTrain
LDA ★	0.90	**0.889**	0.80	1.00	0.940	P5	0.060
SVM	0.85	0.842	0.80	0.90	0.920	P5	0.133
LR Binary	0.85	0.824	0.70	1.00	0.860	P5	0.176
RF	0.85	0.842	0.80	0.90	0.920	P5	0.158
LR Bin. ★	0.90	**0.900**	0.90	0.90	0.950	P7	0.080
SVM	0.85	0.857	0.90	0.80	0.960	P7	0.136
LDA	0.85	0.842	0.80	0.90	0.950	P7	0.132
RF	0.80	0.778	0.70	0.90	0.930	P7	0.222
CART	0.85	0.842	0.80	0.90	0.855	P7	0.095
Naive Bayes	0.75	0.706	0.60	0.90	0.875	P7	0.214

[a] ★ = Best model of the phase (Ph). ΔTrain = F1_train − F1_test. Sens. = Sensitivity. Spec. = Specificity

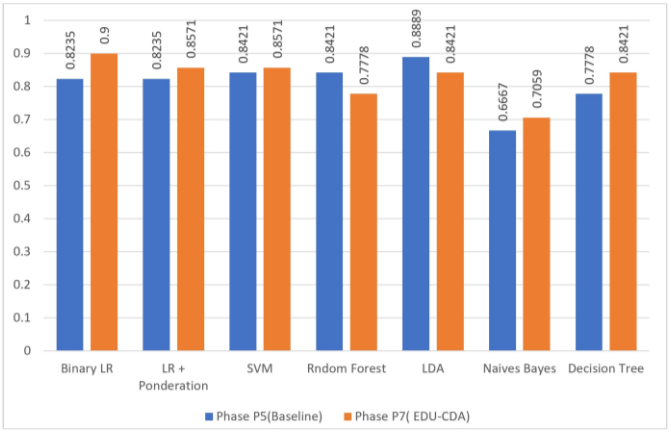

Fig. 4. Comparative F1 performance — Phases P5 (Baseline) and P7 (Post-EDU-CDA) on the real test set – before and after augmentation.

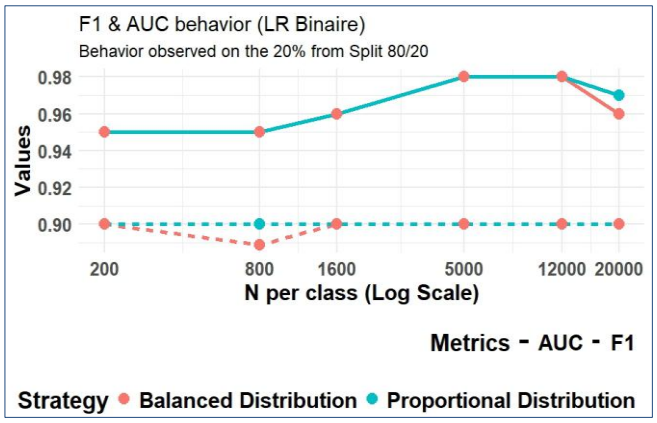

Fig. 5. Analysis of sensitivity to the volume of augmentation.

Three observations emerge from this sensitivity analysis. Initially, F1 plateaus at N = 200 for the proportional strategy (F1 = 0.900 across all configurations) as observed in Fig. 5. The model reaches its maximum binary discrimination capacity very quickly. Next, the AUC follows an upward trajectory up to N = 5,000–12,000 (AUC = 0.98), reflecting a gradual improvement in probability calibration without changing the binary decisions.

In addition, the balanced strategy exhibits a local degradation at N = 800 (F1 = 0.889, Sensitivity = 0.80), suggesting instability in the generation method at this volume level. The proportional strategy is therefore preferred.

D. Cross-Validation and Sensitivity Analysis (RQ2)

5-fold cross-validation on real data (CV1) confirms the hierarchy: LDA is the best model (AUC = 0.944, σ = 0.029), while Binary LR has the lowest AUC (0.810, σ = 0.086). Cross-validation on augmented data (CV2) yields remarkable stability, as shown in Table VII. The increase in AUC from 0.810 to 0.995 provides empirical evidence of the contribution of EDU-CDA.

TABLE VII. CROSS-VALIDATION ON AUGMENTED DATA (BINARY LR, N=800)

Fold	AUC	Sensitivity	Specificity
Fold 1	0.9926	0.9688	0.9688
Fold 2	0.9967	0.9875	0.9813
Fold 3	0.9964	0.9750	0.9750
Fold 4	0.9985	0.9750	0.9875
Fold 5	0.9929	0.9438	0.9938
Average ★	**0.9954**	0.9700	0.9813
Standard deviation	0.0026	0.0162	0.0099

The influence of the volume of synthetic data on the performance of the binary LR model was systematically studied across six levels of N (200 to 20,000 observations per class) and two generation strategies. This evaluation, presented in Table VIII, was conducted exclusively on the 20 real observations in the test set to ensure the external validity of the results.

TABLE VIII. CROSS-VALIDATION FOR EACH DISTRIBUTION AND N

Strategy	N/class	Total obs	AUC	SD
Proportional	200	400	0.9815	0.0153
Proportional	800	1,600	0.9954	0.0026
Proportional	1,600	3,200	0.9940	0.0034
Proportional	5,000	10,000	0.9946	0.0004
Proportional	20,000	40,000	0.9945	0.0008
Balanced	200	479	0.9870	0.0124
Balanced	800	1,679	0.9948	0.0034
Balanced	5,000	10,079	0.9951	0.0009
Balanced	20,000	40,079	0.9944	0.0014

The cross-validation stability analysis based on the increase in sample size shows that N = 200 exhibits significant instability for both strategies, with high standard deviations (0.0153 and 0.0124, respectively), indicating that the model has not yet stabilized at this level. Starting at N = 800, the proportional strategy simultaneously achieves the highest AUC (0.9954) and the lowest SD (0.0026).

N = 800 with the proportional strategy emerges as the optimal configuration: it is the only point where the model combines a maximum AUC, a minimum SD, and stability demonstrated fold by fold, without requiring an excessive volume of data.

As observed in Fig. 6, the best-ranked model is LDA (F1 = 0.831), closely followed by Binary LR (F1 = 0.805). The confidence intervals of the five best models overlap substantially, ruling out any reliable ranking.

```
=== F1 Mean + IC 95% ===
> print(ci, row.names = FALSE)
                Modele    F1 IC_inf IC_sup
                   LDA 0.831  0.726  0.935
             LR Binaire 0.805  0.690  0.921
 LR + Pondération 0.801  0.686  0.916
    Random Forest 0.794  0.666  0.921
                   SVM 0.790  0.672  0.908
    Decision Tree 0.720  0.593  0.847
      Naive Bayes 0.705  0.555  0.855
```

Fig. 6. Confidence interval (screenshot on terminal).

Two consequences follow. CV3 shows that EDU-CDA raises Binary LR, the most penalized model on real data alone (CV1: AUC = 0.810), to statistical parity with the best baseline model (LDA), corroborating the contribution of EDU-CDA (H2) while qualifying the idea of a strict ranking. Fig. 7 shows the results of applying the McNemar test to the out-of-fold predictions.

```
=== McNemar – LDA vs OTHER MODELS ===
> print(do.call(rbind, lapply(setdiff(modeles, tete),
+                             function(m) mcnemar_pair(tete, m))), row.names = FALSE)
             Comparison A_best B_best      p Significant
      LDA vs LR Binaire      3      3 1.0000         Non
LDA vs LR + Pondération      3      3 1.0000         Non
             LDA vs SVM      7      2 0.1797         Non
   LDA vs Random Forest      9      4 0.2668         Non
     LDA vs Naive Bayes      9      1 0.0215         Oui
   LDA vs Decision Tree     15      6 0.0784         Non
```

Fig. 7. McNemar test applied to the out-of-fold predictions.

Binary LR is selected as the final model for two reasons, independent of raw performance. To begin with, it makes no assumption about the distribution of the predictors, whereas LDA assumes multivariate normality within each group — an assumption ill-suited to a predominantly binary dataset (9 variables out of 16). Second, it provides odds ratios together with their confidence intervals, offering a direct and auditable interpretation of the orientation factors that LDA does not allow in such a readable form.

E. Final Model - Optimal Threshold and Odds Ratios (RQ1)

The Youden index identifies the threshold 0.6061 (Sensitivity = Specificity = 0.90) as optimal. Table IX and Fig. 8 present the odds ratios of the final model estimated on the augmented data (N = 800, 95% CI).

TABLE IX. ODDS RATIOS OF THE FINAL BINARY LR MODEL (N=800, 95% CI)

Cariable	OR	IC Inf.	IC Sup.
(Intercept)	52,729	1,887	1698,036
age	1,445	1,303	1,629
style_travail	0,433	0,294	0,632
test_logique	1,648	1,181	2,362
test_creativite	0,107	0,058	0,175
niv_bac_plus	2,521	1,136	5,826
fin_famille	0,477	0,213	1,036
fin_institutionnel	32,399	3,021	657,726
temps_intensif	3,199	1,353	7,979
dom_numerique_tech	3,402	1,546	7,759
dom_autre	0,898	0,329	2,464
lang_java	29,913	3,643	498,063
lang_js	1,123	0,266	6,189
trans_redac	2,552	1,168	5,791
nb_trans	1,402	1,006	1,983
profil_dev	5,628	4,066	8,346
profil_designer	0,178	0,122	0,244

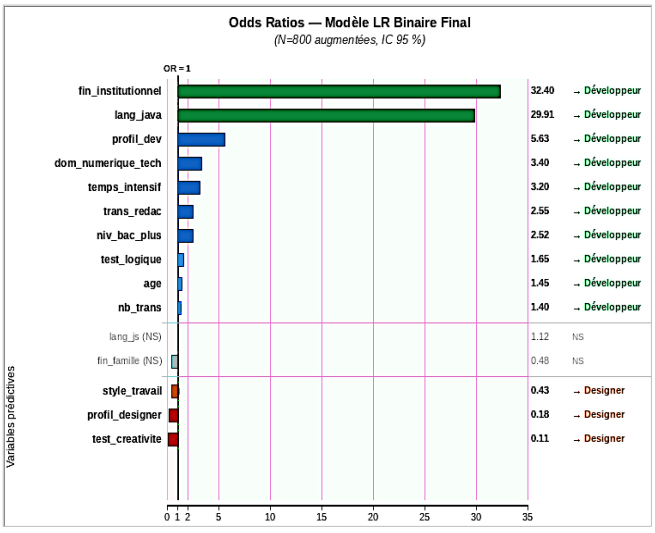

Fig. 8. Odds ratios of the final binary LR model.

F. Class-Wise Analysis and Confusion Matrix (RQ1)

Table X presents the final model's performance by class on the real test set (20 observations). Both classes exhibit symmetrical performance (Precision = Recall = F1 = 0.90), reflecting the quality of the EDU-CDA balancing and the effectiveness of the chosen Youden threshold.

TABLE X. PERFORMANCE BY CLASS VIA BINARY LR AND POST-EDU-CDA (P7) ON THE REAL TEST SET

Class	Precision	Recall	F1-score	Support
Graphic Designer	0.90	0.90	0.90	10
Application Developer	0.90	0.90	0.90	10
Macro average	**0.90**	**0.90**	**0.90**	20
Overall accuracy	—	—	**0.90**	20

The Binary LR after EDU-CDA is the only model to exhibit a perfectly symmetrical error distribution (1 FN, 1 FP), confirming the sensitivity/specificity balance (0.90/0.90) observed in the aggregated metrics. Table XI presents this performance.

TABLE XI. PERFORMANCE BY CLASS BY MODEL

Model	VP	FN	FP	VN
LR Binary ★	9	1	1	9
SVM	9	1	2	8
LR + Weighting	9	1	2	8
LDA	8	2	1	9
Decision Tree	8	2	1	9
Random Forest	7	3	1	9
Naive Bayes	6	4	1	9

After boosting, Binary LR is the only model to achieve F1 = 0.900 on the test set while maintaining the lowest training-test gap among competitive models (0.080). Random Forest continues to exhibit persistent overfitting (gap = 0.222). Binary LR is selected as the production model. Table XII presents the confusion matrix.

TABLE XII. CONFUSION MATRIX BINARY LR POST-EDU-CDA (P7)

Model	F1 Training	F1 Test	Gap
Random Forest	1.000	0.778	0.222
SVM	0.993	0.857	0.136
LDA	0.974	0.842	0.132
Binary LR	0.980	0.900	0.080

G. Final Learn-Orient Classification Model (RQ4)

Applying the Youden index to the ROC curve of the final model yields two thresholds at J = 0.80. The graph in Fig. 9 illustrates the ROC curve of the final model. It has an AUC of 0.95, indicating excellent overall discriminatory power. It deviates significantly from the reference diagonal (AUC = 0.50 corresponding to a random classifier), confirming that the model effectively distinguishes between Designer and Developer profiles across all possible thresholds.

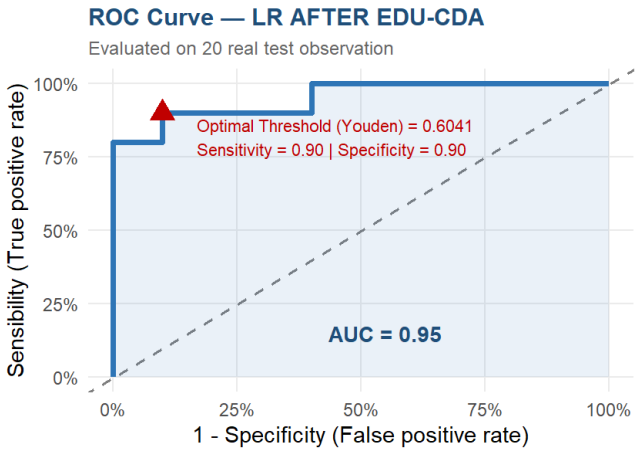

Fig. 9. ROC Curve of the final model after EDU-CDA applied.

The threshold of 0.6061 is selected as the optimal threshold. When two thresholds achieve the same Youden index, statistical convention recommends selecting the lower one, which in this case corresponds to a perfect symmetrical balance between sensitivity and specificity (0.90/0.90). Presented in Table XIII, this balance is particularly well-suited to the context of career guidance, where the two types of error, incorrectly directing a Developer toward Design and incorrectly directing a Designer toward Development, have comparable consequences for the candidate.

TABLE XIII. OPERATIONAL TESTS - LEARN-ORIENT SYSTEM ON 3 PROFILES

Profile Tested	$\bar{d}$ Gower	% threshold τ	Prediction	Confidence
Test 1 — Actual designer	0.092	15%	Designer 99.9%	High
Test 2 — Actual developer	0.175	28%	Developer 100.0%	High
Test 3 — Atypical profile	0.377	60%	Developer 100.0%	Moderate
Gower threshold $\tau = 0.629$	—	100%	—	—

The real profiles (Tests 1 and 2) have Gower distances less than 28% of the threshold $\tau = 0.629$, with clearly defined probabilities and High confidence. The artificial atypical profile (Test 3) correctly triggers a Moderate confidence level at 60% of the threshold, answering RQ4 in the affirmative.

V. DISCUSSION

A. Discriminability of Pre-Training Profiles (RQ1)

Learn-Orient achieves F1 = 0.900 and AUC = 0.950 on the real test set without standardized academic variables, answering RQ1 positively with performance well above the operational threshold. This result is even more notable as it is obtained on only 99 observations and surpasses the ceiling observed by Al-Dossari et al. [12] (70.47% with XGBoost on a dataset twenty times larger). It confirms that the performance of a guidance system does not depend simply on the volume of data, but on the relevance of the variables and the quality of the augmentation method.

Two counterintuitive predictors deserve attention. Institutional funding (OR = 32.40) outperforms all technical skills: this is not an intrinsic characteristic of the candidate but a characteristic of their context of access to training. This strong signal suggests that government scholarship programs structurally attract technical profiles, targeting fields perceived as strategic (software development). This result opens important perspectives for policies on funding retraining.

The creativity test score (OR = 0.107) can be interpreted considering Holland's Person-Environment Fit theory [38]: self-reported creativity is a stable indicator of fit with the Graphic Designer profile, regardless of formal technical skills. Recent studies in the French-speaking context of career transition [39] confirm that these person–environment fit factors remain robust determinants of career change decisions. Song et al. [17] similarly emphasize the importance of motivational and behavioral variables beyond academic characteristics alone.

B. Effectiveness of the EDU-CDA Method (RQ2)

The Binary Logistic Regression model test on the test set of 20 observations yields a F1 = 0.900. This represents a meaningful indicator of the validity of the EDU-CDA Method. It can be explained by a well-documented principle [10] which establishes that the simplest models, those most vulnerable to small samples, benefit the most from augmentation. Before EDU-CDA, quasi-linear separation on 79 observations produced a perfect training score (F1_train = 1.000). After augmentation to 1,600 observations, this phenomenon disappears (F1_train = 0.980), and the model generalizes correctly.

EDU-CDA addresses a specific shortcoming of standard SMOTE [11] : its inapplicability to constrained binary variables. The TSTR of 0.924, obtained on a dataset twenty times smaller and predominantly binary than that of Chinodakufa et al. [27] (TSTR = 0.997 on 10,000 continuous observations), falls within defensible values. The results of Anggrawan et al. [21] and Cu et al. [22] show similar stabilization patterns from equivalent augmentation volumes.

CV3 cross-validation applies the augmentation within each fold and tests on real data. It confirms that EDU-CDA moves Binary LR from the status of the most vulnerable model to statistical parity with the best baseline model (LDA): across the 99 real observations, Binary LR and LDA do not differ significantly in any case (McNemar test, p = 1.000). The realistic generalization performance of Binary LR stands at around F1 = 0.80 [0.69; 0.92], the value of 0.900 obtained on the single test set reflecting a favourable composition of the split. EDU-CDA, therefore, does improve Binary LR.

C. Fidelity of Synthetic Data and the Gower Filter (RQ3 and RQ4)

The three-part protocol fully validates RQ3. The only TVD value slightly exceeding 0.05 (end_family, TVD = 0.061) concerns the variable whose highest proportion (65.7% family funding) makes exact reproduction via conditional Bernoulli more sensitive to random fluctuations, an expected behavior with no practical impact on overall quality. The combination of these three dimensions into a single protocol had not been formalized in the literature on educational data.

Operational tests validate RQ4: the Gower filter correctly distinguishes well-covered profiles (distances < 28% of the threshold) from atypical profiles (60% of the threshold). The IQR normalization proposed by D'Orazio [37] is essential: without it, continuous variables with high interquartile ranges would dominate the calculation, mechanically reducing the contribution of binary variables.

D. Limitations

While the proposed model and methodology demonstrate good performance, some limitations are acknowledged:

L1- Representativeness: the sample obtained does not claim to be statistically representative of all individuals undergoing digital retraining in Benin. The validity of the study rests on the relevance and diversity of the collected profiles rather than on their representativeness in the probabilistic sense. The results and the developed system constitute a proof of concept, and a decision-support tool based on the patterns observed in this sample; generalizing these findings to other contexts will require external validation on new cohorts. CV3 cross-validation also shows that the five best models are statistically indistinguishable at this sample size. The F1 differences reported in phase P7 should therefore be read as indicative, and confirmation on a larger sample remains desirable.

L2 — Statistical power: With 20 observations in the test set, a misclassified observation accounts for 5 F1 points. Confirmation is therefore required on a larger sample.

L3 — Sample bias: overrepresentation of males (71.9%) and near-exclusivity of Beninese individuals (95.6%) amplified in the synthetic data.

L4 — Partial conditional independence: dependencies between binary variables not covered by the two consistency rules are not preserved by the Bernoulli component.

L5 — Self-reporting bias: technical skills are not objectively verified, constituting a source of unmeasurable noise.

L6 — Restriction to two fields: Audiovisual Specialist (n=11) and Digital Project Manager (n=8) remain outside the scope due to insufficient sample sizes.

L7 — Gower threshold calculated on synthetic data: the threshold τ characterizes the internal dispersion of the augmented space. External validation on new cohorts is necessary.

L8 — Lack of hyperparameter optimization: the algorithms are evaluated using their default parameters. Optimization could alter the performance hierarchy for the SVM and RF.

L9 — Discriminant variable selection criteria: Selecting variables based on three criteria could exclude important variables with a small number of observations.

VI. CONCLUSION

Career guidance toward digital professions in Sub-Saharan Africa remains, to date, a largely subjective, non-scalable, and poorly equipped process. This study builds on this concrete observation to propose a response that is both scientific and operational, grounded in real data from different programs and training centers or universities in Benin.

The first contribution of this study (Learn-Orient) demonstrates that it is possible to statistically distinguish between two career transition profiles based exclusively on pre-training characteristics: without grades, without formal assessment, and without structured interviews. With an F1 score of 0.900 and an AUC of 0.950 on real-world data, the system generates interpretable probabilistic recommendations, accompanied by a graded confidence level that explicitly signals its own limitations. The Gower filter ensures responsible use by identifying atypical profiles. This last point, beyond being a technical detail, serves as a genuine ethical safeguard. A system that knows when not to respond is more useful and more responsible than a system that always responds.

The second contribution (EDU-CDA) addresses a specific methodological challenge that the literature had not yet resolved: how to augment educational data characterized by constrained binary mixed variables, low volume, and imbalanced classes? The proposed solution is simple in its logic, rigorous in its implementation, and validated across three independent dimensions. It transformed logistic regression, the most vulnerable model on raw data, into the most stable model after augmentation, with an AUC rising from 0.810 to 0.995 in cross-validation. This result suggests that simple, interpretable models are not inherently inferior to complex models; they generally require more data to realize their full potential.

Beyond the metrics, the results yield operational insights that decision-makers can directly apply. The fact that institutional funding is the strongest predictor of the Developer profile (OR = 32.40), ahead of all technical skills, is not a statistical anomaly. This would suggest that training in this field is costly and that support in this area is likely the best way to provide high-quality, large-scale training for this profession. It also signals that funding policies are already, unintentionally, shaping the profiles of those entering certain fields. Similarly, the creativity score as the primary predictor of the Designer profile (OR = 0.107) confirms that cognitive abilities measured before training are robust indicators of career orientation, regardless of prior academic background.

In the context of using Learn-Orient, these results call for four concrete actions: institutionalize pre-admission profiling using the 16 identified variables; differentiate funding mechanisms according to the targeted fields of study; use the atypical profile signal as a trigger for enhanced human support; and actively diversify the pool of candidates to correct current demographic biases.

This study is a first step, not an endpoint. Extending the system to other fields, such as Audiovisual Specialist, Digital Project Manager, Data Analyst, and others, as soon as staffing allows, would broaden the system's coverage. Integrating a component that estimates the probability of success conditional on the field of study would transform Learn-Orient into a true recommendation tool rather than merely a profile similarity tool. External validation on new training cohorts, followed by deployment as a web application accessible to educational advisors, constitute the two immediate operational steps. Finally, the EDU-CDA method is designed to be transferable without major modification to similar programs, particularly in Africa. It is worth noting that it requires neither a large-scale data collection infrastructure nor expertise in deep generative models.

In a context where 230 million digital jobs are expected in sub-Saharan Africa by 2030 [5], properly guiding each candidate from the outset is not a methodological luxury. It is a direct lever for employability, equitable access, and the efficiency of resources invested in training.

Ethical Considerations

All data was collected with the informed consent of the respondents. The data is anonymized: no personally identifiable information is used in the analyses, and the results are presented exclusively in aggregated form. Learn-Orient is designed as a decision-support tool complementary to human assessment, not as an autonomous decision-making system. The educational advisor retains full responsibility for the final guidance. The profile recognition filter ensures that the system flags its own limitations for atypical profiles. In accordance with Rudin's [15] recommendations for systems with high human stakes, the odds ratios from the logistic regression allow the counselor to explain the reasons behind the generated recommendation to the candidate with the utmost transparency.

Acknowledgment

Our sincere thanks go to Firmin Monyevedo TOVODOUNNON, CEO of EIG Group, Issiakou SOULEYMANE, CEO of the School of Digital Professions (École des Métiers du Numérique - EMN) in Benin, all other training centers and schools that have shared their data with us, as well as all the candidates who agreed to participate in this study. This work is part of the research activities of the Computer Science Research and Decision Support Laboratory (L@RIAD), from the Institute of Mathematics and Physical Sciences (IMSP), an entity of the University of Abomey-Calavi (UAC), whose supervisory team we would like to thank.

References

[1] R. A. Marte Espinal et L. Fabián V, « Determinantes de la deserción universitaria: un estudio de caso en la República Dominicana », SIJIS, vol. 2, no 1, p. 255-268, mars 2021, doi: 10.51798/sijis.v2i1.76.

[2] G. Wydra-Somaggio, « Early termination of vocational training: dropout or stopout? », Empirical Res Voc Ed Train, vol. 13, no 1, p. 5, déc. 2021, doi: 10.1186/s40461-021-00109-z.

[3] International Labour Office, Global Employment Trends for Youth 2022: investing in transforming futures for young people. S.l.: INTL LABOUR OFFICE, 2022.

[4] OECD, Dir., OECD Digital Economy Outlook 2024 (Volume 2): Strengthening Connectivity, Innovation and Trust. dans OECD Digital Economy Outlook. Paris: OECD Publishing, 2024. doi: 10.1787/3adf705b-en.

[5] IFC - World Bank, Empowering Africa's youth: Bridging the digital skills gap. 2121 Pennsylvania Avenue, NW Washington, DC 20433 USA: World Bank, 2024.

[6] M. S. Macharia, « Reengineering mass career acquisition through technical vocational education training counseling in Kenya », IJRBS, vol. 8, no 6, p. 212-218, oct. 2019, doi: 10.20525/ijrbs.v8i6.533.

[7] F. Eicker, G. Haseloff, et B. Lennartz, Vocational Education and Training in Sub-Saharan Africa: Current Situation and Development. Bielefeld: W. Bertelsmann Verlag, 20170316.

[8] G. Adankon et P. Houngue, « How Can Intelligent Educational Technologies Address the Challenges of Professional Retraining in Africa: From Current State to Future Perspectives? », dans 2025 IEEE

Global Engineering Education Conference (EDUCON), London, United Kingdom: IEEE, avr. 2025, p. 1-8. doi: 10.1109/EDUCON62633.2025.11016638.

[9] F. Trujillo, M. Pozo, et G. Suntaxi, « Artificial intelligence in education: A systematic literature review of machine learning approaches in student career prediction », J. Technol. Sci. Educ., vol. 15, no 1, p. 162, mars 2025, doi: 10.3926/jotse.3124.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, et W. P. Kegelmeyer, « SMOTE: Synthetic Minority Over-sampling Technique », jair, vol. 16, p. 321-357, juin 2002, doi: 10.1613/jair.953.

[11] Haibo He et E. A. Garcia, « Learning from Imbalanced Data », IEEE Trans. Knowl. Data Eng., vol. 21, no 9, p. 1263-1284, sept. 2009, doi: 10.1109/TKDE.2008.239.

[12] H. Al-Dossari, F. A. Nughaymish, Z. Al-Qahtani, M. Alkahlifah, et A. Alqahtani, « A Machine Learning Approach to Career Path Choice for Information Technology Graduates », Eng. Technol. Appl. Sci. Res., vol. 10, no 6, p. 6589-6596, déc. 2020, doi: 10.48084/etasr.3821.

[13] P. Houngue, M. Hountondji, et T. Dagba, « An Effective Decision-Making Support for Student Academic Path Selection using Machine Learning », IJACSA, vol. 13, no 11, 2022, doi: 10.14569/IJACSA.2022.0131184.

[14] M. Wu, G. Subramaniam, D. Zhu, C. Li, H. Ding, et Y. Zhang, « Using Machine Learning-based Algorithms to Predict Academic Performance - A Systematic Literature Review », dans 2024 4th International Conference on Innovative Practices in Technology and Management (ICIPTM), Noida, India: IEEE, févr. 2024, p. 1-8. doi: 10.1109/ICIPTM59628.2024.10563566.

[15] C. Rudin, « Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead », Nat Mach Intell, vol. 1, no 5, p. 206-215, mai 2019, doi: 10.1038/s42256-019-0048-x.

[16] D. W. Hosmer, S. Lemeshow, et R. X. Sturdivant, Applied Logistic Regression, 1re éd. dans Wiley Series in Probability and Statistics. Wiley, 2013. doi: 10.1002/9781118548387.

[17] Q. C. Song, H. J. Shin, C. Tang, A. Hanna, et T. Behrend, « Investigating machine learning's capacity to enhance the prediction of career choices », Personnel Psychology, vol. 77, no 2, p. 295-319, juin 2024, doi: 10.1111/peps.12529.

[18] L. K. Smirani, H. A. Yamani, L. J. Menzli, et J. A. Boulahia, « Using Ensemble Learning Algorithms to Predict Student Failure and Enabling Customized Educational Paths », Scientific Programming, vol. 2022, p. 1-15, avr. 2022, doi: 10.1155/2022/3805235.

[19] J.-E. Lee, A. Jindal, S. N. Patki, A. Gurung, R. Norum, et E. Ottmar, « A comparison of machine learning algorithms for predicting student performance in an online mathematics game », Interactive Learning Environments, vol. 32, no 9, p. 5302-5316, oct. 2024, doi: 10.1080/10494820.2023.2212726.

[20] F. E. Arévalo-Cordovilla et M. Peña, « Comparative Analysis of Machine Learning Models for Predicting Student Success in Online Programming Courses: A Study Based on LMS Data and External Factors », Mathematics, vol. 12, no 20, p. 3272, oct. 2024, doi: 10.3390/math12203272.

[21] Information Technology Education Department, Bumigora University, Indonesia, A. Anggrawan, H. Hairani, et C. Satria, « Improving SVM Classification Performance on Unbalanced Student Graduation Time Data Using SMOTE », IJIET, vol. 13, no 2, p. 289-295, 2023, doi: 10.18178/ijiet.2023.13.2.1806.

[22] N. G. Cu, T. L. Nghiem, T. H. Ngo, M. T. L. Nguyen, et H. Q. Phung, « Increment of Academic Performance Prediction of At‐Risk Student by Dealing With Data Imbalance Problem », Applied Computational Intelligence and Soft Computing, vol. 2024, no 1, p. 4795606, janv. 2024, doi: 10.1155/2024/4795606.

[23] C. Esteban, S. L. Hyland, et G. Rätsch, « Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs », 4 décembre 2017, arXiv: arXiv:1706.02633. doi: 10.48550/arXiv.1706.02633.

[24] M. Hernandez, P. A. Osorio-Marulanda, M. Catalina, L. Loinaz, G. Epelde, et N. Aginako, « Comprehensive evaluation framework for synthetic tabular data in health: fidelity, utility and privacy analysis of generative models with and without privacy guarantees », Front. Digit. Health, vol. 7, p. 1576290, avr. 2025, doi: 10.3389/fdgth.2025.1576290.

[25] J. Jordon, A. Wilson, et M. van der Schaar, « Synthetic Data: Opening the data floodgates to enable faster, more directed development of machine learning methods », 2020, arXiv. doi: 10.48550/ARXIV.2012.04580.

[26] Z. Qian, B.-C. Cebere, et M. van der Schaar, « Synthcity: facilitating innovative use cases of synthetic data in different data modalities », 18 janvier 2023, arXiv: arXiv:2301.07573. doi: 10.48550/arXiv.2301.07573.

[27] T. A. Chinodakufa, A. A. Shafin, et K. M. Ahmed, « Synthetic Data in Education: Empirical Insights from Traditional Resampling and Deep Generative Models », 2026, arXiv. doi: 10.48550/ARXIV.2604.21031.

[28] G. E. A. P. A. Batista et M. C. Monard, « An analysis of four missing data treatment methods for supervised learning », Applied Artificial Intelligence, vol. 17, no 5-6, p. 519-533, mai 2003, doi: 10.1080/713827181.

[29] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, et O. Tabona, « A survey on missing data in machine learning », J Big Data, vol. 8, no 1, p. 140, oct. 2021, doi: 10.1186/s40537-021-00516-9.

[30] I. Guyon et A. Elisseeff, « An Introduction to Feature Extraction », dans Feature Extraction, vol. 207, I. Guyon, M. Nikravesh, S. Gunn, et L. A. Zadeh, Dir., dans Studies in Fuzziness and Soft Computing, vol. 207. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, p. 1-25. doi: 10.1007/978-3-540-35488-8_1.

[31] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 0 éd. Routledge, 2013. doi: 10.4324/9780203771587.

[32] C. F. Dormann et al., « Collinearity: a review of methods to deal with it and a simulation study evaluating their performance », Ecography, vol. 36, no 1, p. 27-46, janv. 2013, doi: 10.1111/j.1600-0587.2012.07348.x.

[33] N. Japkowicz et M. Shah, Evaluating learning algorithms: a classification perspective, First paperback ed. New York: Cambridge University Press, 2014.

[34] J. A. Hanley et B. J. McNeil, « The meaning and use of the area under a receiver operating characteristic (ROC) curve. », Radiology, vol. 143, no 1, p. 29-36, avr. 1982, doi: 10.1148/radiology.143.1.7063747.

[35] W. J. Youden, « Index for rating diagnostic tests », Cancer, vol. 3, no 1, p. 32-35, 1950, doi: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

[36] J. C. Gower, « A General Coefficient of Similarity and Some of Its Properties », Biometrics, vol. 27, no 4, p. 857, déc. 1971, doi: 10.2307/2528823.

[37] [37] M. D'Orazio, « Distances with mixed type variables some modified Gower's coefficients », 2021, arXiv. doi: 10.48550/ARXIV.2101.02481.

[38] J. L. Holland, Making vocational choices: a theory of vocational personalities and work environments, 3rd ed. Odessa, Fla: Psychological Assessment Resources, 1997.

[39] Yassine JAOUI and Mouna HILMI, « les facteurs explicatifs de la reconversion professionnelle : Proposition d'un modèle conceptuel » [Explanatory factors of professional retraining: Proposal of a conceptual model], jul. 2024, doi: 10.5281/ZENODO.12805028. [In French]

[40] L. A. Goodman, « Snowball Sampling », Ann. Math. Statist., vol. 32, no 1, p. 148-170, mars 1961, doi: 10.1214/aoms/1177705148.

[41] M. Q. Patton, Qualitative research & evaluation methods, 3. ed., [Nachdr.]. Thousand Oaks: Sage, 2010.

[42] H. Gumuchian and C. Marois, Initiation à la recherche en géographie: Aménagement, développement territorial, environnement [Introduction to geographical research: Planning, territorial development, environment]. Montréal: Presses de l'Université de Montréal, 2000. doi: 10.4000/books.pum.14790. [In French]

[43] K. Jordan, C. Myers, K. Damani, P. Khagame, A. Mumbi, et L. Njuguna, « Supporting equitable access to learning via SMS in Kenya: Impact on engagement and learning outcomes », Brit J Educational Tech, vol. 56, no 4, p. 1530-1552, juill. 2025, doi: 10.1111/bjet.13533.

[44] T. Hastie, R. Tibshirani, et J. Friedman, The Elements of Statistical Learning. dans Springer Series in Statistics. New York, NY: Springer New York, 2009. doi: 10.1007/978-0-387-84858-7.