

DOI: <https://doi.org/10.24297/ijct.v25i.9795>

Modelling the Employability of Management Graduates: Complementing Parametric Approaches with Machine Learning on Small Social Data

Ravelonahina Andrianjaka Hasina¹, Robinson Matio², Andriamanohisoa Hery Zo³

¹ Doctoral Student in The Doctoral School of Science and Engineering and Technic Innovation (STII): Laboratory of Cognitive Science and Application. University of Antananarivo, Madagascar.

² Doctor HDR (Habilitation to Supervise Research) in STII, University of Antananarivo.

³ Professor in STII, University of Antananarivo.

Abstract

This study investigates how supervised and unsupervised machine learning algorithms can complement traditional statistical methods in the analysis of social survey data. Social science datasets are typically small, noisy, and heterogeneous, which makes robustness and interpretability more important than computational efficiency.

Using data from a 2024 survey on the employability of management graduates in Antananarivo, the study compares machine learning approaches with classical multivariate techniques. The objectives are to provide a statistical description of a social reality and to establish criteria for selecting algorithms suited to small-sample contexts.

The methodological framework integrates statistical tools such as Chi-square tests, analysis of variance, and multiple regression with exploratory approaches including association rules and clustering. It also incorporates supervised models such as neural networks trained via gradient descent and its variants. Beyond these models, ensemble methods based on decision trees—bagging, random forests, and gradient boosting—are evaluated to highlight their relative strengths.

Findings show that gradient boosting offers the most consistent predictive performance while remaining relatively simple to implement. This makes it particularly effective for analysing small and heterogeneous datasets, thereby providing practical value for applied social science research.

Keywords: Multivariate statistics, machine learning, gradient descent, boosting.

1 Introduction

In social surveys, datasets are often small compared to those used in large-scale data mining. The main challenge in applying data mining or machine learning techniques to such data is not computational efficiency, but rather the selection of algorithms that remain appropriate and robust. Methods must handle variability, limit overfitting, and provide reliable results despite heterogeneous and noisy samples.

In Madagascar, many young graduates face difficulties in entering the labour market due to a mismatch between the skills acquired during their education and the evolving requirements of employers. This situation is particularly acute in management-related fields, where specialized and up-to-date qualifications are increasingly demanded. Identifying the factors that influence employability is therefore essential to address this gap.

Traditional multivariate analysis methods are widely applied in the social sciences, but they rely on restrictive assumptions such as linearity, normality, and homoscedasticity. In contrast, machine learning models infer relationships directly from the data, tolerate violations of these assumptions, and incorporate mechanisms such as regularization, variable selection, and feature importance estimation. They can also manage mixed data types and remain robust in the presence of noise or missing values.

In this study, we apply a range of statistical and machine learning techniques to analyse employability data from young Malagasy graduates. While neural networks represent a powerful modelling approach, their implementation in this context is challenging due to the limited sample size, the heterogeneity of variables, and the risk of overfitting. These constraints reduce their practical applicability in social survey research. By contrast, decision tree-based ensemble methods, and particularly gradient boosting, prove more adaptable to the data structure and yield consistent and interpretable results, making them especially well-suited for this type of analysis.

2 Methodology and Models Employed

The questionnaire, written in French and composed primarily of closed-ended questions, was administered via Google Forms to respondents holding senior or mid-level management positions. The survey was designed to explore the expectations and recruitment criteria of corporate executives, focusing on five main dimensions: respondent profiles, selection criteria, desired skills, changes in academic programs, and the evaluation of universities. The data, collected

primarily in Antananarivo, June 2024, comprise 214 individuals and are structured around 60 mixed variables (both quantitative and qualitative). The analysis draws on both traditional multivariate methods and machine learning algorithms for classification tasks (Krzywinski & Altman, 2017): Bagging, Random forest (Breiman, 2001a), the Apriori algorithm for association rule mining, the k -Nearest Neighbors (k -NN) algorithm, Support Vector Machine (SVM), and gradient descent (GD) variants (Bottou, 1991). Another category of algorithms that has demonstrated strong performance is ensemble methods for aggregating weak learners, such as boosting (Mayr et al., 2014). All these methods were applied to uncover meaningful patterns and partitions within our dataset.

https://github.com/kope67/Machine_Learning_vs_Parametric_Approaches_on_Small_Social_Data

2.1 Algorithmic Approaches Implemented

Classical parametric approaches, such as linear regression, rely on strong assumptions regarding data distribution and the functional form of relationships between variables. While they are interpretable and effective on well-structured datasets, their applicability is limited in high-dimensional contexts, especially with heterogeneous, noisy, or partially missing data. In our study, the mixed nature of the 60 variables and the complexity of interactions between graduate characteristics and recruiter criteria make these methods insufficient to capture complex or nonlinear structures.

« As a result, we found that the most famous DM techniques in our context are: SVM, Naïve Bayes, PCA, Logistic Regression, k -means algorithm, KNN, decision tree, Neural networks, text mining and Item Response Theory. From this Systematic Literature Review (SLR), from 2005 to 2019 and according to the accuracy, we conclude that logistic regression, decision trees, ANN, Random Forest, Text-IRT are the best DM techniques for employability studies » (Moumen et al., 2020).

This combined approach enables the assessment of model robustness, the detection of latent data structures (Breiman, 2001b), and the extraction of predictive profiles associated with employability.

The analyses were primarily conducted using Python, which provided the main environment for machine learning, statistical modelling, and performance evaluation. R and Tanagra were employed in a complementary role, mainly to generate selected tables, figures, and exploratory outputs, thereby enhancing the interpretability and presentation of the results.

2.2 The multiple linear regression model:

$D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ be n pairs of random variables defining a sample. The possible values are in $X \times Y$ where $X \subset \mathbb{R}^d$ and $Y = \mathbb{R}$ (regression).

For a classical linear regression, we minimize the sum of squared errors:

$$L(\beta) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (1)$$

$$\text{with } \hat{y}_i = X_i \beta$$

We have the decomposition:

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 \quad (2)$$

The coefficient of determination is defined as:

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} \quad (3)$$

2.3 Non-parametric and unsupervised methods

Unlike parametric approaches, non-parametric methods do not assume a predefined functional form for the relationships between variables. Their complexity automatically adapts to the data structure, providing great flexibility, especially in exploratory contexts such as employability analysis.

2.3.1 Unsupervised clustering: K-means

Algorithms such as K -means were used to identify groups of recruiter or graduate profiles, revealing trends not visible through factor analysis alone.

The K-means algorithm is one of the most commonly used clustering techniques. It partitions a dataset into K clusters, where K is a user-specified number. The goal of K-means is to minimize the sum of distances between each data point and the centroid of the cluster to which it is assigned. Association Rule Mining:

2.3.2 The Apriori algorithm

The Apriori algorithm highlighted frequent relationships between skills, experiences, and recruitment criteria. Support is defined as the probability:

$$\text{Support}(X) = \text{Pr}(X)$$

This corresponds to the proportion of transactions containing item X within the set of all transactions.

$$\text{Support}(X \cap Y) = \text{Pr}(X \cap Y)$$

The proportion of transactions containing both item X and item Y within the set of all transactions.

Confiance:

$$\text{Pr}(Y|X) = \frac{\text{Support}(X \Rightarrow Y)}{\text{Support}(X)} = \frac{\text{Pr}(X \cap Y)}{\text{Pr}(X)} \quad (4)$$

Lift:

$$\text{lift}(X \Rightarrow Y) = \frac{\text{Support}(X \Rightarrow Y)}{\text{Support}(X) \times \text{Support}(Y)} = \frac{\text{Pr}(X \cap Y)}{\text{Pr}(X) \times \text{Pr}(Y)} \quad (5)$$

2.4 Supervised methods

$D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ n pairs of random variables defining a sample.

The possible values are $X \times Y$ where $X \subset R^d$ and $Y = \{-1; 1\}$ (classification) or $Y = R$ (regression).

We aim to solve a regression problem

$$Y_i = F^*(X_i) + \varepsilon_i \quad (6)$$

With $F^*: R^d \rightarrow R$ measurable and ε_i i.i.d such that $E[\varepsilon_i | X_i] = 0$.

2.4.1 Heuristic:

The objective is to find a function between the variable X and the variable Y; $F^*(X) = Y$.

A deterministic relationship may not necessarily exist, yet the goal is to identify the best F^* that models the interdependence.

For a chosen function class F (shallow decision trees, linear regression, shallow neural networks, naïve classifiers, etc.), such an F^* is theoretically defined by:

$$F^* = E[L(Y, F(X))] \quad (7)$$

Where $L: R \times R \rightarrow R$ is a convex and integrable loss function. The previous solution is then approximated by an empirical version

$$F^* = \sum_{i=1}^n L(Y_i, F(X_i)) \quad (8)$$

In our study, several supervised algorithms were tested.

2.4.2 GD, SGD, and Mini-Batch GD.

We began with classical approaches such as Gradient Descent (GD), Stochastic Gradient Descent (SGD), and Mini-batch Gradient Descent.

Initially, we trained **a single neuron using logistic regression** with gradient descent. Our Python implementation details all steps, including coefficient initialization, sigmoid activation, gradient computation, coefficient updates, and optimization to minimize the log-loss function.

$D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ n pairs of random variables defining a sample. The possible values are $X \times Y$ where $X \subset R^d$ and $Y = \{-1; 1\}$ (classification)

- The sigmoid activation function is defined as:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{9}$$

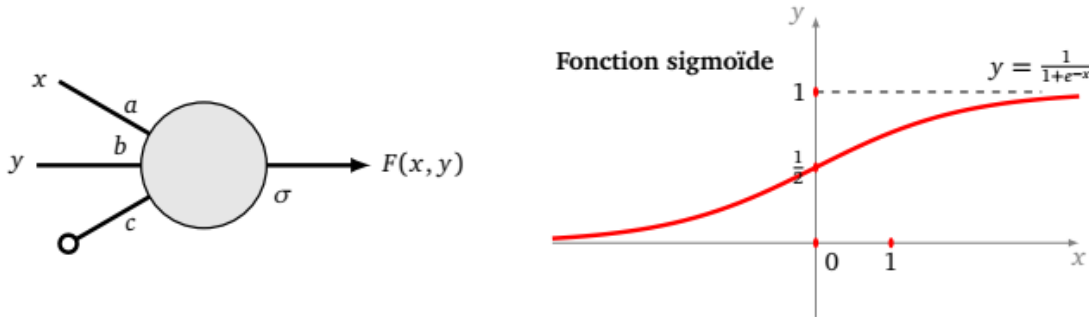


Figure 1: A neuron with two inputs equipped with the sigmoid activation function

$$F(x; y) = \sigma(ax + by + c)$$

More generally $F(x_i) = \sigma(\theta^T x_i)$ pour $x_i \in \mathbb{R}^d$

- The data are split into training and test sets, and the weights are adjusted using gradient descent algorithms: standard (GD), stochastic (SGD), or mini-batch.
- The loss function used is the log-loss, commonly employed in binary classification, which is based on the principle of likelihood maximization within a probabilistic framework.

$$L(\omega, b) = -\frac{1}{n} \sum_{i=1}^n \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \tag{10}$$

The initial learning model is **logistic regression, equivalent to a single-neuron neural network** with a sigmoid activation function. It is a binary classification model taking multiple quantitative variables as input.

$$\theta_j^{(t+1)} = \theta_j^{(t)} - \eta \cdot \frac{1}{n} \sum_{i=1}^n (\sigma(\theta^T x_i) - y_i) x_i^{(j)} \tag{11}$$

2.4.3 Support Vector Machines (SVM)

Kernel Support Vector Machines (SVMs) are powerful classifiers that handle nonlinear data by mapping inputs into higher-dimensional spaces via a kernel function (Takeda et al., 2006). Kernel theory is central to SVMs and enables the solution of complex classification problems.

Advantages:

- High-dimensional efficiency:** Effective when the number of features exceeds the number of samples.
- Nonlinear class separation:** Kernels allow modelling complex relationships between classes.
- Strong generalization:** The maximum-margin principle helps maintain good performance on test data.

2.5 Decision trees

2.5.1 Methodological Perspectives in Machine Learning with trees

Building on this analysis, several methodological avenues can be explored to improve both the performance and interpretability of models applied to complex field data. These perspectives aim to fully integrate recent advances in machine learning into employability analysis while ensuring rigorous interpretation of results to inform training and recruitment policies.

$X \subset \mathbb{R}^d$, $x \in X$ and $D_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ be n data pairs defining a sample.

$x_i \in R^d$ and $y_i \in R$ (regression) or $y_i \in \{0; 1\}$ (classification).

- A tree is a single predictor $x \mapsto h(x; a)$, piecewise constant and obtained via recursive dyadic partitioning. The parameter a of a tree $h(\cdot; a)$ characterizes the split points of the data space X into L regions: ($L \geq 1$) and $(R_l)_{l=1}^L$ corresponding to the leaves (terminal nodes).

- If $\bigcup_{i=1}^n E_i$ forms a partition of the universe, the average amount of information required to randomly identify an event E_i is given by **Shannon entropy**, defined as:

$$- \sum_{i=1}^n p_i \times \log_2 p_i \tag{12}$$

Entropy decreases measure the reduction of disorder provided by a splitting criterion, enabling a decision tree to better separate classes at each node.

2.5.2 CART Decision Trees (Breiman et al., 2017)

At each step, a node is split into two child nodes.

- Resampling of D_n into D_n' is performed via bootstrap (with possible repetition) by drawing $\in \llbracket 1; n \rrbracket$ samples (hyperparameters to be set initially), with $|D_n'| = t \leq n$.
- The split is made along one of the d possible coordinates, ϵ_{dir} being the set of possible splitting coordinates.
- The algorithm stops when the minimum number of observations per cell is reached (hyperparameters to be set in advance).

Let $A \subset X$, $|A|$ denotes its cardinality, i.e., the number of observations $X_i = (X_i^{(1)}, \dots, X_i^{(d)})$.

Let $z \in [0; 1]$ and $j \in \llbracket 1; d \rrbracket$ be a possible splitting direction in ϵ_{dir} .

We define the set of candidate splits as:

$$\mathcal{E}_{cut} = \{(j; z) \in \llbracket 1; d \rrbracket \times [0; 1] / j \in \epsilon_{dir}\} \tag{13}$$

For each pair $(j; z)$, the CART criterion is:

$$L(j; z) = \frac{1}{|A|} \sum_{i=1}^n (Y_i - \bar{Y}_A)^2 - \frac{1}{|A|} \sum_{i=1}^n \left(Y_i - \bar{Y}_{AL} 1_{\{x_i^{(j)} < z\}} - \bar{Y}_{AR} 1_{\{x_i^{(j)} \geq z\}} \right)^2 \tag{14}$$

which represents a measure of the difference between groups

$$A_L = \{x \in A / x^{(j)} < z\} \text{ and } A_R = \{x \in A / x^{(j)} \geq z\}$$

The total dispersion minus the residual dispersions, representing a measure of factorial dispersion, serves as an indicator of the difference between the two groups in ANOVA.

$$(j^*, z^*) \in \arg \arg L(j; z)$$

2.6 Tree Ensembles for Balancing Bias and Variance

2.6.1 Control of the bias-variance trade-off

One of the major challenges in machine learning is the trade-off between bias (systematic errors) and variance (sensitivity to the data). A complex model can reduce bias but increase variance, thereby compromising its ability to generalize.

$$Variance = E \left[\left(\hat{f}(x) - E[\hat{f}(x)] \right)^2 \right] \tag{15}$$

$$Bias^2 = \left(E[\hat{f}(x)] - f(x) \right)^2$$

The expected Mean Squared Error (MSE) can be decomposed into bias, variance, and noise:

$$MSE = E[(y - \hat{f}(x))^2] = \text{Variance} + \text{Bias} + \text{Noise} \quad (16)$$

2.6.2 The random forest

The random forest algorithm reduces variance by combining multiple trees built on random subsamples of data and features. This introduces some bias but stabilizes predictions. Its main advantage lies in this well-controlled bias-variance trade-off.

2.7 Gradient Boosting and CART decision trees

Boosting sequentially combines weak learners (often shallow trees) to correct the errors of the previous model. This approach tends to reduce bias but may increase variance. Gradient Boosting, particularly in its stochastic version (Friedman, 2002), introduces random subsampling to mitigate this risk. This mechanism enhances robustness by limiting overfitting.

We consider the following optimization problem:

$$\{ \min_{F \in \mathcal{F}} L(F) \quad \text{Subject to the constraint that } F \in H$$

The minimization problem can be solved using the standard gradient descent method, i.e., by searching for the function $F \in \mathcal{F}$ that optimizes $L(F)$.

$$F_{m+1} = F_m - \beta \nabla L(F_m) \text{ et } \beta > 0, \quad (17)$$

Knowing that each function F_k is represented by an element of R^n namely $(F_k(x_1), \dots, F_k(x_n))$.

With the constraint $F \in H$ this equality becomes:

$$F_{m+1} = F_m + \beta h_m \quad (18)$$

2.8 Boosting with decision trees (Treeboosting)

$$H = \{\text{Predictor tree with } L \text{ nodes}\}$$

This is the case for the Gradient Boosting Regressor algorithm from the *sklearn.ensemble* module in Python.

$$F_m(x) = F_{m-1}(x) + \sum_{l=1}^L \gamma_{l,m} \mathbf{1}_{\{x \in R_{l,m}\}} \quad (19)$$

$$\gamma_{l,m} = \beta_m \bar{y}_{l,m}$$

Thus, for each region, only one indicator function is nonzero.

$$\sum_{l=1}^L \gamma_{l,m} \mathbf{1}_{\{x \in R_{l,m}\}} = \gamma \quad (20)$$

$$\gamma_{l,m} = \sum_{i=1}^n L(y_i, F_{m-1}(x_i) + \gamma) \quad (21)$$

2.9 Decoupling between variable selection and model specification,

A key methodological challenge is separating the selection of relevant variables from the model's functional form, the nature of variable interactions and their relationship with the target (linear or nonlinear, with or without interactions). This enables more flexible and robust modelling, addressing issues such as multicollinearity and substitution effects between correlated variables.

Lasso regression addresses this by adding a penalty equal to the sum of the absolute values of the coefficients:

$$L_{Lasso}(\beta) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^d |\beta_j| \tag{22}$$

Ridge regression adds a penalty on the L^2 norm of the coefficients:

$$L_{Ridge}(\beta) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^d \beta_j^2 \tag{23}$$

where λ is the regularization parameter (controlling the strength of the penalty) that penalizes the norm of β

3 Results and Discussion

3.1 Statistical approaches implemented

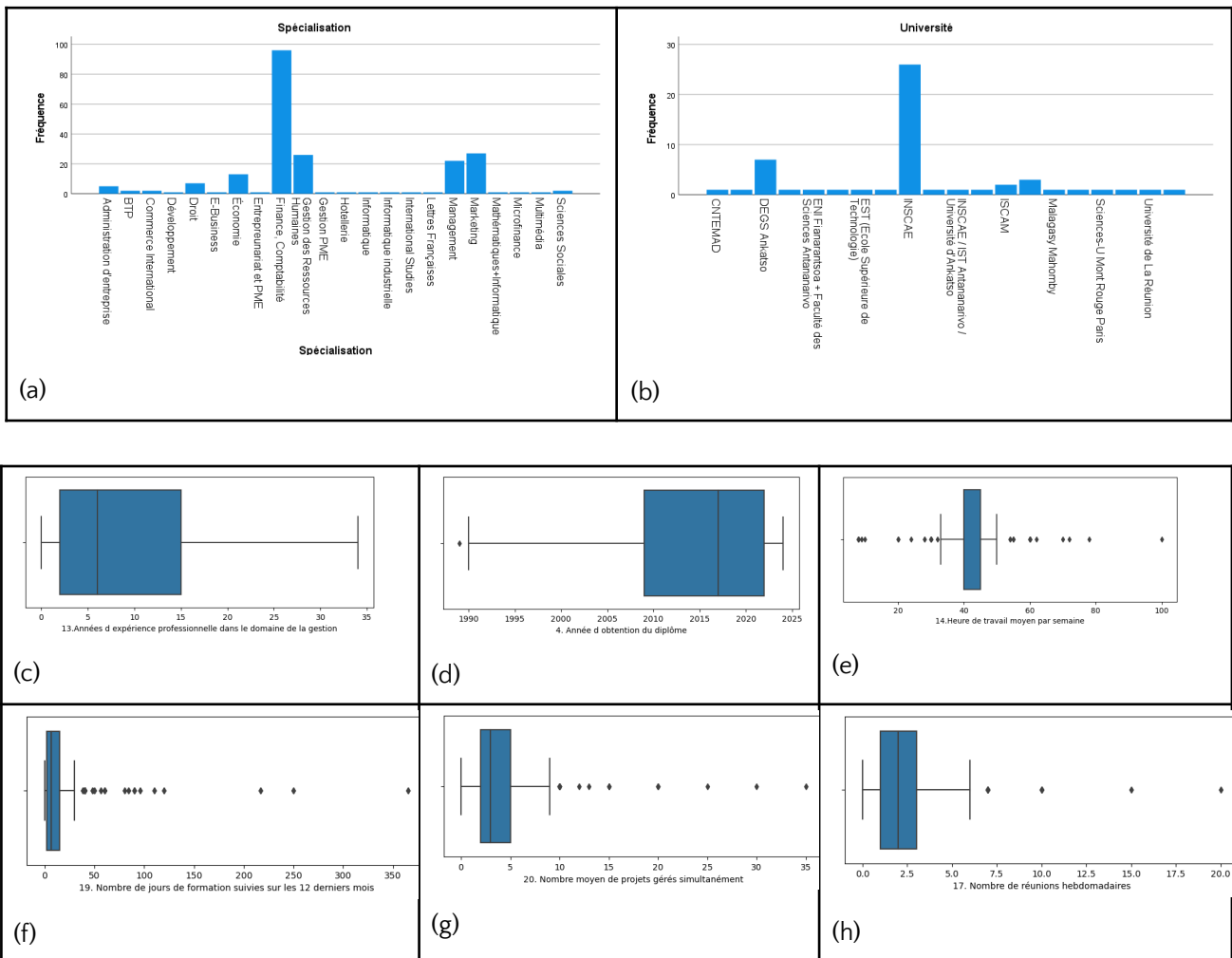


Figure 2: Educational Pathway and Its Implications for Employability

Surveyed executives come from diverse backgrounds, mostly from INSCAE and the University of Ankatso, holding standard managerial positions. While professional experience varies, other characteristics (weekly working hours, annual training days, projects managed, weekly meetings) are relatively homogeneous.

3.1.1 Example of multiple linear regression:

Here, the dependent variable y is the **average number of working hours**, and the explanatory variables are seven other variables listed below. We obtain a linear regression with the following coefficients, along with tests performed on each variable to assess their significance.

Table 1: Regression Results.

Out[120]:	Coef.	Std.Err.	t	P> t	[0.025	0.975]	Multiple Regression	Linear
const	415.686712	214.396588	1.938868	0.054245	-7.666071	839.039494	R ² = 0.082, Adjusted R ² = 0.042, F(7, 163) = 2.075, p = 0.049. N = 171 observations.	
4. Année d'obtention du diplôme	-0.186398	0.106077	-1.757205	0.080760	-0.395859	0.023063		
13. Années d'expérience professionnelle dans le domaine de la gestion	-0.097342	0.126803	-0.767667	0.443795	-0.347730	0.153045		
15. Taille de votre équipe actuel	-0.006515	0.005136	-1.268406	0.206462	-0.016657	0.003627		
16. Taille de l'équipe la plus grande que vous avez gérée	0.003235	0.004613	0.701215	0.484169	-0.005875	0.012344		
17. Nombre de réunions hebdomadaires	0.768499	0.297059	2.587027	0.010554	0.181919	1.355079		
19. Nombre de jours de formation suivies sur les 12 derniers mois	0.017772	0.019275	0.922019	0.357881	-0.020289	0.055833		
20. Nombre moyen de projets gérés simultanément	0.019864	0.145439	0.136579	0.891532	-0.267323	0.307050		

In our example, the coefficient of determination is $R^2 = 0.082$ indicating that the linear model is clearly not appropriate.

Data context:

- **Total:** 214 observations.
- **Split:** 80% for training the model (training set) and 20% for evaluation (test set).
- **Model:** multiple linear regression estimated using *statsmodels*.

Metric Train Test Comment

R²	0.0818	-0.042	The R ² on the training set (≈ 8.2 %) is very low, indicating that the model explains almost none of the variability of the target variable. The negative R ² on the test set shows that the model predicts worse than the mean on new data.
MSE	76.15	277.79	The mean squared error on the test set is extremely high, confirming poor generalization.
RMSE	8.73	16.67	The mean error (in the units of the target variable) nearly doubles from the training to the test set, indicating a substantial gap between learning and generalization.

$$R^2_{test} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{24}$$

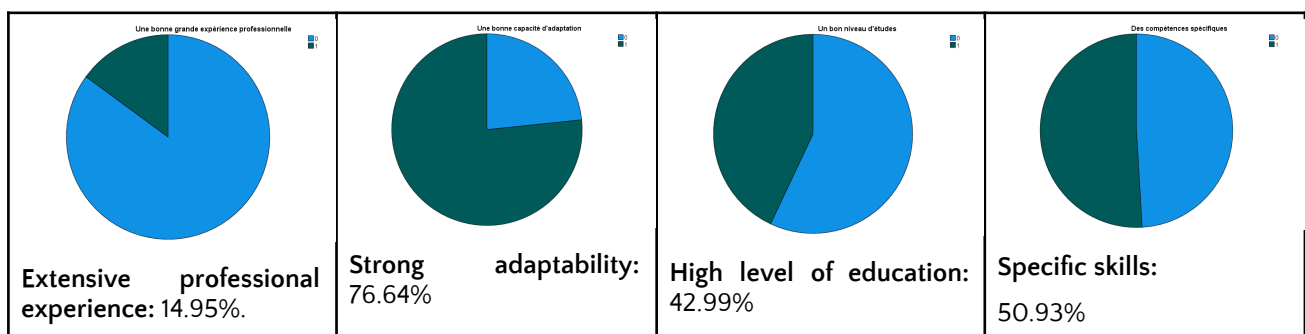
A negative R² on the test set is a clear signal that the model fails to generalize, indicating a need to reconsider either the variable selection or the type of model.

Probable causes:

- **Underfitting:** the model is too simple (high bias) and fails to capture any meaningful relationships.
- **Irrelevant variables:** there is a lot of noise in the explanatory data.
- **Nonlinear relationships:** the linearity imposed by the regression does not match the true relationship.

3.1.2 Main expectations of recruiters

In our survey, executives were asked to indicate their primary expectations of young graduates during the recruitment process. Respondents were allowed to select multiple answers. The results **highlight the key skills and qualities that recruiters prioritize**, providing insight into the competencies considered most relevant in the labour market for management graduates. Respondents could select multiple options among the four proposed.



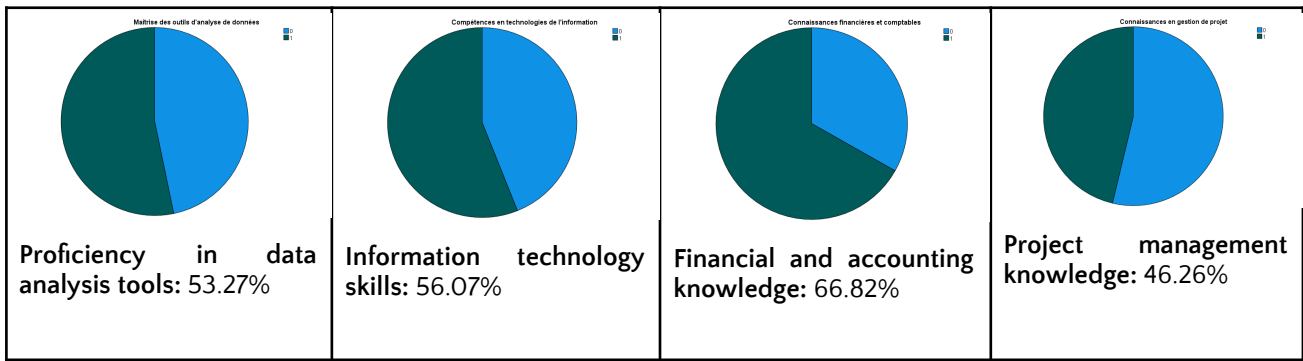


Figure 3: Essential Technical Skills for a Recent Graduate in Management

For specific positions or certain departments within the company, there is a tendency to favor certain universities (a). Among the executives we approached, 42.99% indicated that their preference for certain universities depends on the department or position, while 25.23% expressed clear specific preferences. The remaining respondents reported either having no specific preferences (23.36%) or valuing diversity in institutions (8.41%).

However, giving preference to candidates from foreign universities is not systematic (b).

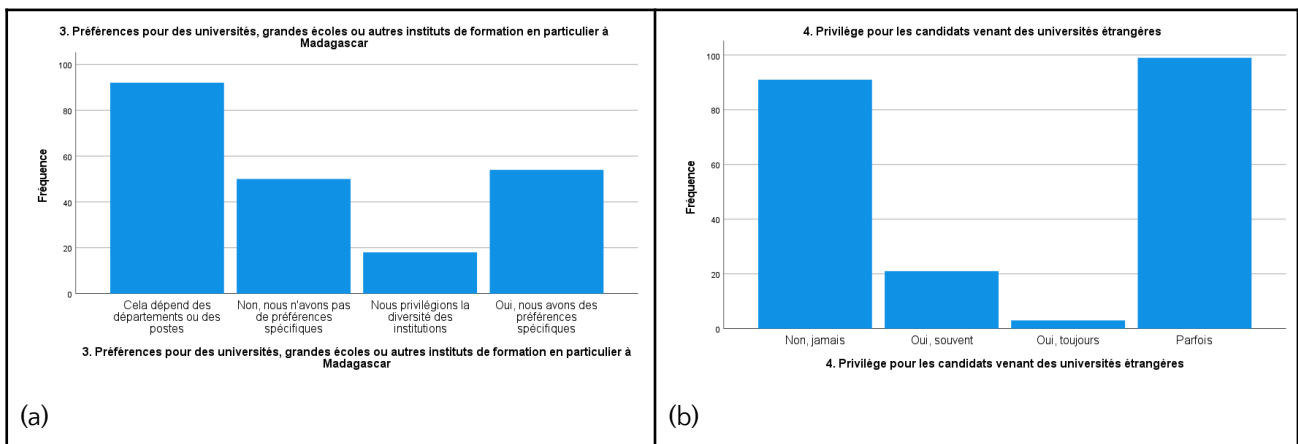


Figure 4: Preference for Universities

Let us test the hypothesis of independence between salary and training participation.

Table 2: Chi-squared tests

Diplôme le plus élevé obtenu	Remunération mensuelle actuelle (approximativement)
<p>Graphique à barres</p> <p>Suivi formation après diplôme</p> <p>■ Non (Blue)</p> <p>■ Oui (Green)</p> <p>Effectif</p> <p>Diplôme le plus élevé obtenu</p>	<p>Graphique à barres</p> <p>Remunération mensuelle actuelle (approximativement)</p> <p>■ Entre 2 Millions MGA et 3 Millions MGA (Blue)</p> <p>■ Moins de 2 Millions MGA (Green)</p> <p>■ Plus de 3 Millions MGA (Red)</p> <p>Effectif</p> <p>Suivi des formations complémentaires après l'obtention du diplôme</p>
<p>Chi-squared test</p> <p>$p = 0.600$ (ns).</p> <p>21 cells (80.8%) with an expected count < 5 (min = 0.40)</p>	<p>Chi-squared test</p> <p>$\chi^2 = 0.705$, $df = 2$, $p = 0.703$ (ns).</p> <p>No cells with an expected count < 5 (min = 21.45)</p>

We observe that the test is not significant, indicating that "participation in additional training" is independent of salary.

The relationship with the level of degree could not be assessed because there are too many cells with expected

counts below 5, requiring some regrouping.

This variable, "participation in additional training," will serve as the primary variable in our predictive algorithms.

Analysis of Variance:

Table 3: Influence of Perceived IT and Collaborative Platform Importance on Programming Skills Assessment

Attribute_Y		Attribute_X		Description				Statistical test		
Importance de la maîtrise des technologies de l'information (ERP, CRM) et des plateformes collaboratives (Microsoft Teams, Slack)		Importance du savoir-programmer (Python, R, SQL) pour un jeune diplômé en gestion		Value	Examples	Average	Std-dev	Variance decomposition		
				Peu important	84	3,1548	0,7362	Source	Sum of square	d.f.
				Important	95	3,2842	0,5954	BSS	1,0020	3
				Très important	18	3,3333	0,8402	WSS	99,8438	210
				Pas important	17	3,2941	0,7717	TSS	100,8458	213
				All	214	3,2383	0,6881	Significance level		
				Statistics	Value	Proba				
				Fisher's F	0,702477	0,551527				

Computation time : 0 ms.
Created at 27/08/2025 10:24:42

Anova: F = 0,702 $df_1 = 3$, $df_2 = 210$, p = 0.551 (ns).

The results (Table 3) show that, although a slight difference in means was observed across groups, this difference was not statistically significant. In other words, the perceived importance of programming is not meaningfully affected by the importance attributed to information technologies and collaborative platforms.

3.1.3 Linear Correlation Coefficient and Regression Model

Initially, we considered only quantitative variables to examine potential relationships between them.

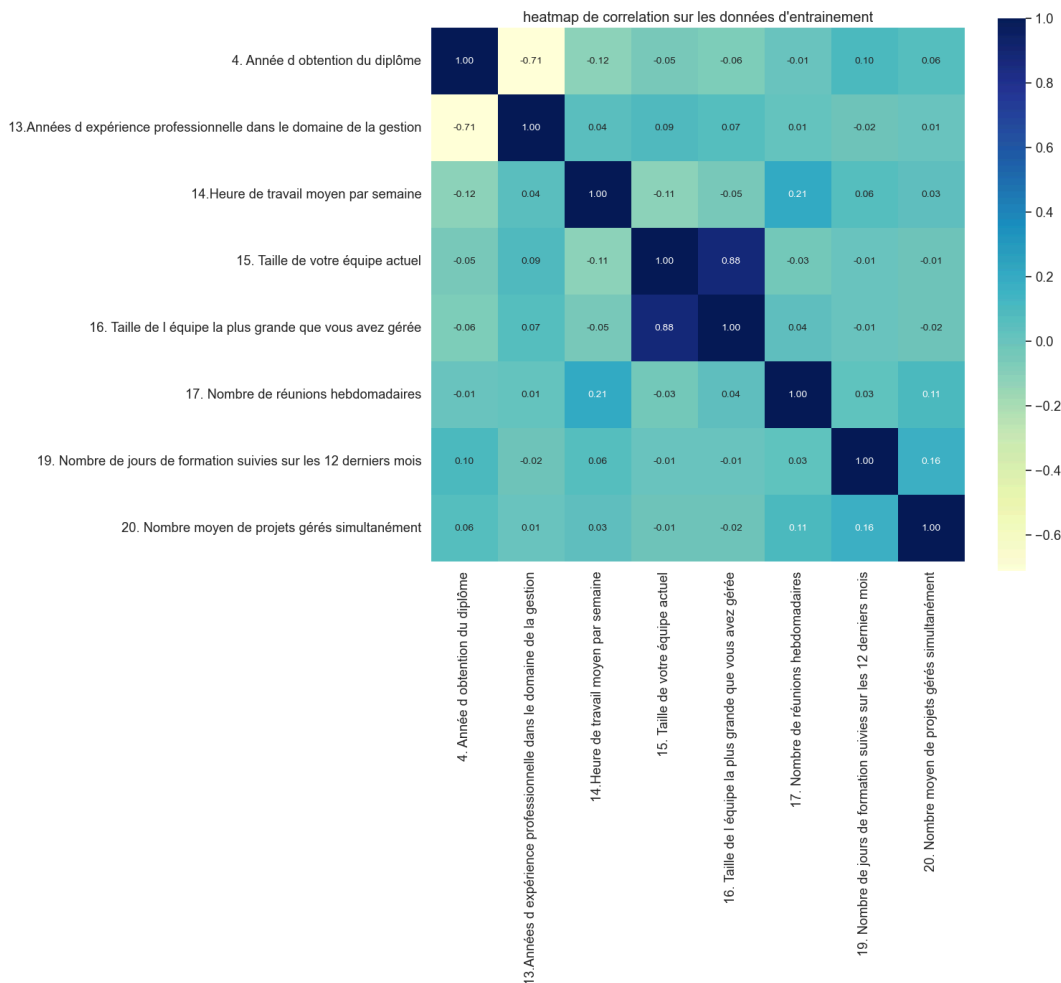


Figure 5: Linear Correlation Coefficients

We observed that the correlations are generally rather weak as shown in Figure 5.

3.1.4 Clustering of manager types

Table 4: Manager types

K-means clustering with 3 clusters of sizes 10, 3, 201					
Cluster means:					
	Expérience_professionnelle Nombre_moyen_projets	Heure_par_semaine	Taille_équipe_actuel	Taille_équipe_grande	
1	11.050000	38.50000	271.70000	524.50000	3.400000
2	11.666667	38.33333	2033.33333	2166.66667	4.666667
3	9.152985	41.32836	15.50746	23.07463	4.064677

In this example, managers were grouped into three clusters using the K-means algorithm, based on the following five variables. Managers of larger teams generally have more experience, while those with smaller teams tend to work longer hours as presented in Table 4.

3.1.5 Frequent relationships

We aimed to identify the most frequently occurring associations to better understand company's expectations. For this, we used association rules, specifically the Apriori algorithm, with a minimum support of 0.33 and a minimum confidence of 0.75.

For example: We obtained a common association

« A strong educational background » ⇒ « Good adaptability skills »

What about the required personal qualities and attitudes?

Table 5: Frequent relationships

N°	Antecedent(s)	Consequent	Support (%)	Confidence (%)	Lift	Occurrences
1	Proactivity, Adaptability	Teamwork, Critical Thinking	35,5	86,4	1,31	76
2	Proactivity, Thinking	Teamwork, Critical Adaptability	35,5	80,9	1,24	76
3	Proactivity, Thinking	Adaptability, Critical Teamwork	35,5	89,4	1,23	76
4	Teamwork, Critical Thinking	Adaptability	40,7	79,8	1,22	87
5	Teamwork, Adaptability	Critical Thinking	40,7	79,8	1,21	87
6	Proactivity, Teamwork	Critical Thinking	43,9	78,3	1,19	94
7	Adaptability, Critical Thinking	Teamwork	40,7	85,3	1,18	87

N° Antecedent(s)	Consequent	Support (%)	Confidence (%)	Lift	Occurrences
8 Proactivity, Adaptability	Critical Thinking	39,7	77,3	1,17	85
9 Teamwork, Adaptability, Critical Thinking	Proactivity	35,5	87,4	1,13	76
10 Proactivity, Critical Thinking	Teamwork	43,9	81,0	1,12	94
11 Teamwork, Critical Thinking	Proactivity	43,9	86,2	1,11	94
12 Proactivity, Adaptability	Teamwork	41,1	80,0	1,10	88

The test yielded twelve possible associations (Table 5), among which the one with the highest support and confidence is:

'Teamwork and Collaboration'; 'Critical Thinking and Problem Solving' ⇒ 'Proactivity and Initiative Taking'.

This algorithm allows the detection of frequent relationships, such as associations between variables like *general knowledge of AI*, *AI adoption in one's sector*, *company readiness to integrate AI*, and *trust in AI*, while optimizing the process through the use of frequency and confidence thresholds.

Table 6: The Apriori algorithm on AI-related data.

No. Antecedents (lhs)	Consequent (rhs)	Support	Confidence	Coverage	Lift	Count
1 Company readiness to integrate AI = Moderately ready	= Trust in AI = Moderately confident	0.3131	0.8072	0.3879	1.3187	67
2 General knowledge of AI = "Machines learn from data to make predictions or decisions"; Company readiness to integrate AI = Moderately ready	= Trust in AI = Moderately confident	0.2150	0.7931	0.2710	1.2956	46
3 AI already adopted in your sector? = Moderately adopted	= Trust in AI = Moderately confident	0.2804	0.7059	0.3972	1.1531	60

The integration of AI in companies poses considerable challenges, particularly regarding trust, data management, and ethical responsibility.

3.2 Experimental Results with GD, SGD, and Mini-Batch GD

In this study, we sought to predict whether an individual *participates in supplementary training* ("formation_complement") by leveraging two variables: *professional experience* ("13.exp_pro") and *team size* ("15.Taille_équipe").

Using a standard Gradient Descent (GD) algorithm, we obtained an *accuracy* (proportion of correct predictions) of 0.63.

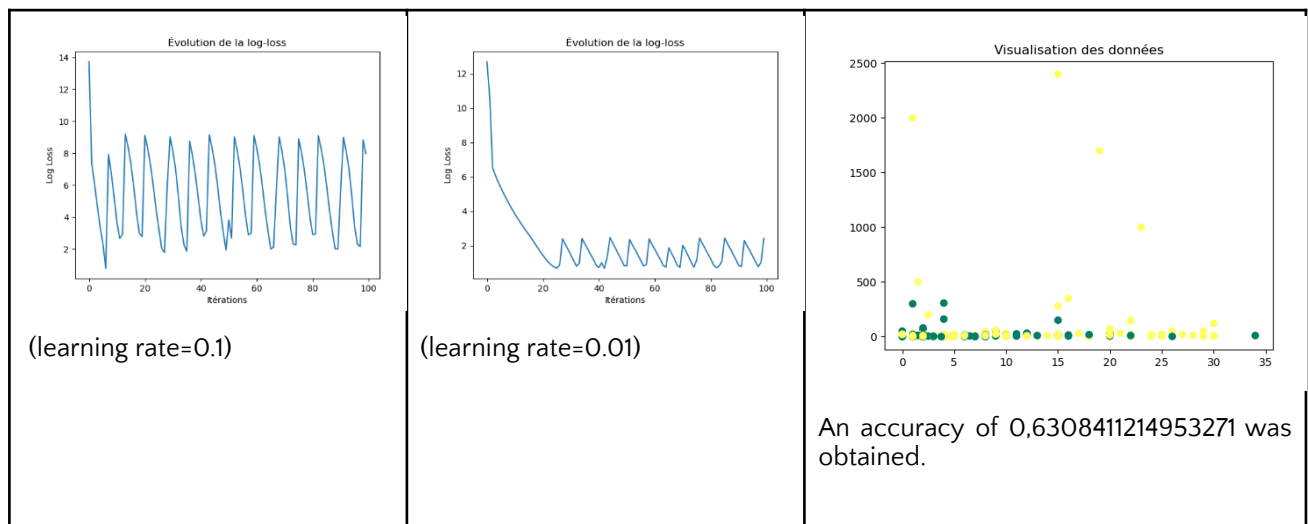


Figure 6: Classification Results obtained using standard Gradient Descent.

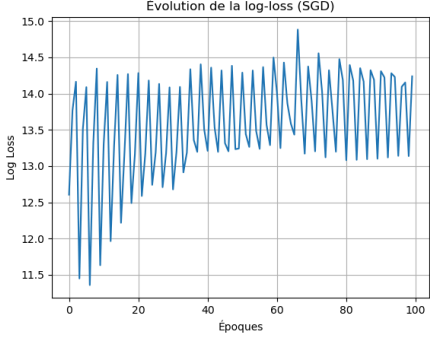
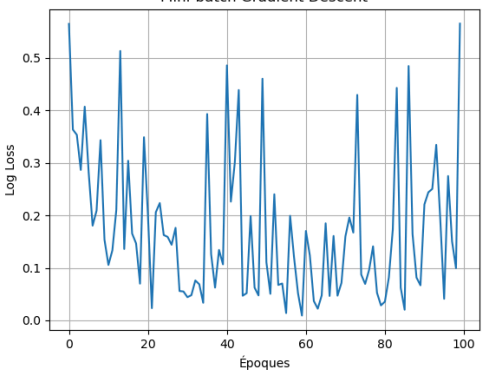
Here (as shown in Fig. 6), the curve shows strong instability in the log-loss: instead of decreasing steadily or stabilizing at a low value, it oscillates noticeably at each iteration.

Comment on stability:

- The model does not reach a stable convergence.
- The repeated oscillations indicate that the loss function fails to find a clear minimum, instead cycling between states.

Possible explanations:

1. **Learning rate too high**
→ The gradient descends too quickly, “overshooting” the minimum at each update.
2. **Highly noisy data**
→ Random variations in the data lead to inconsistent updates of the loss function.
3. **Unregularized model**
→ Without adequate penalization, the model may overfit random fluctuations, maintaining instability.
4. **Poor initialization**
→ Instability may result from extreme parameter initialization or incorrectly scaled data.

 <p>Figure 7: Obtained using stochastic gradient descent (SGD)</p>	<p>Using a Stochastic Gradient Descent (SGD) algorithm (learning rate=0.1), trained on the entire dataset for 100 epochs, we obtained an accuracy of 0.5187.</p> <p>A prediction probability of 0.52 is barely better than random chance.</p> <p>Possible causes:</p> <ul style="list-style-type: none"> • Underfitted or overly simple model • Poor preprocessing: features not normalized <p>With a learning rate =0.01, we obtain an accuracy of 0.593</p>
 <p>Figure 8: Obtained using mini-batch gradient descent (mini-batch GD)</p>	<p>Accuracy: 0.53</p> <p>We decided to incorporate a larger set of variables, using eight features to predict the target variable 'supplementary training'. The dataset was split into training and test subsets. The independent variables (features) were standardized, and the Mini-Batch Gradient Descent algorithm was applied. The resulting test accuracy was 0.53. Adjustments to the learning rate and the number of iterations did not lead to any significant improvement in test accuracy</p>

3.2.1 Results of Support Vector Machines (SVM)

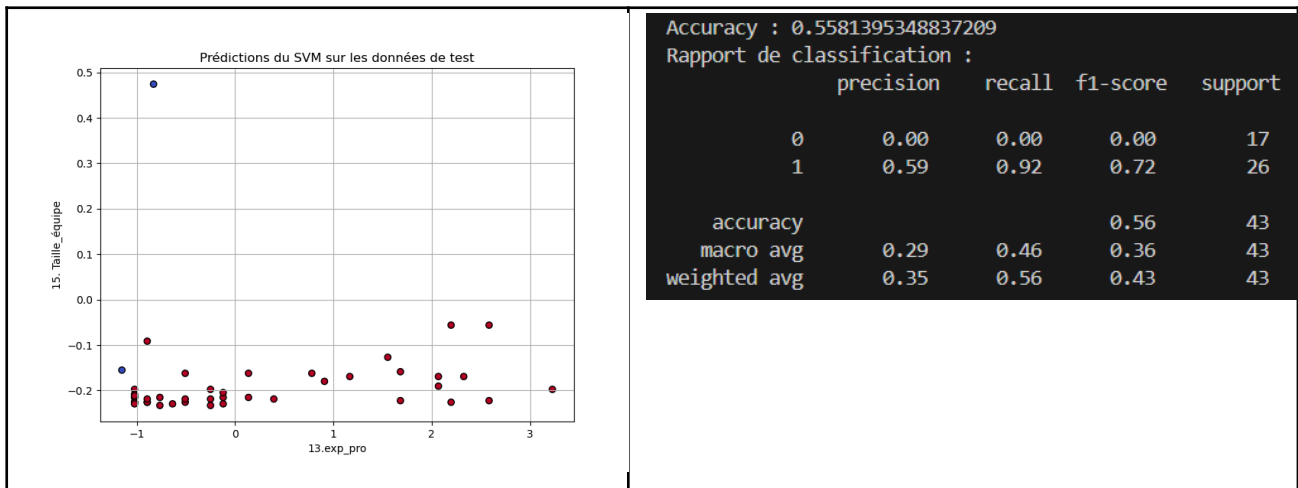


Figure 9: Results of Support Vector Machines (SVM)

In this SVM model, the linear kernel was replaced by the Gaussian RBF (Radial Basis Function) kernel

$$K(x, x') = \exp(-\gamma \|x - x'\|^2)$$

The resulting prediction, with an accuracy of **0.56**, is only marginally better than random guessing. Decision trees

To predict salary: Build a decision tree model to predict a **manager's salary** and identify the most influential features in this prediction. These techniques use measures such as entropy, the Gini index, or other impurity metrics to guide tree construction and determine the best splits and decisions.

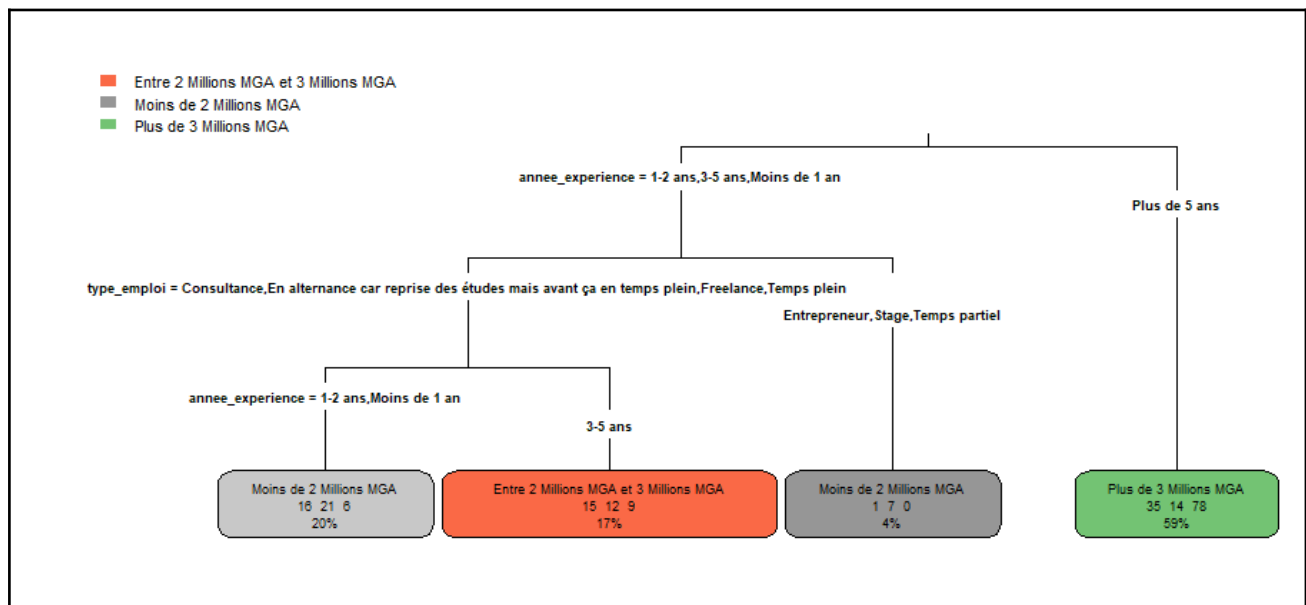


Figure 10: Manager's salary

3.2.2 Experimental Results with CART as classification tree:

$$h(x, a, D_n) = \sum_{l=1}^L c_l 1_{\{x \in R_l\}} \tag{25}$$

with c_l The majority class in the region R_l .

A Python program based on scikit-learn was used to compare impurity criteria for CART:

$$Gini = 1 - \sum_{i=1}^n p_i^2 \tag{26}$$

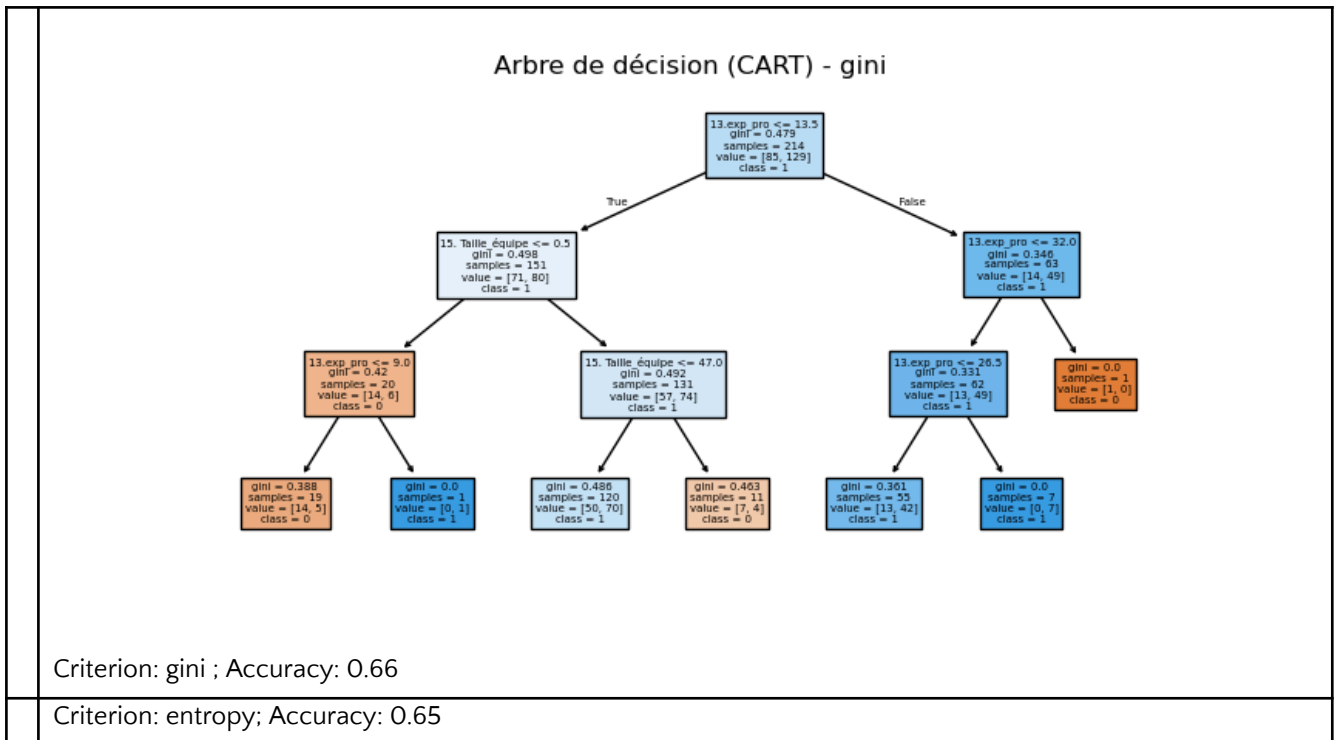


Figure 11: Classification trees

Two decision trees were built to predict participation in supplementary training (Yes/No). Using CART with professional experience and team size as predictors yielded a noticeable accuracy improvement (0.66), though still below a satisfactory threshold.

3.2.3 A regression tree:

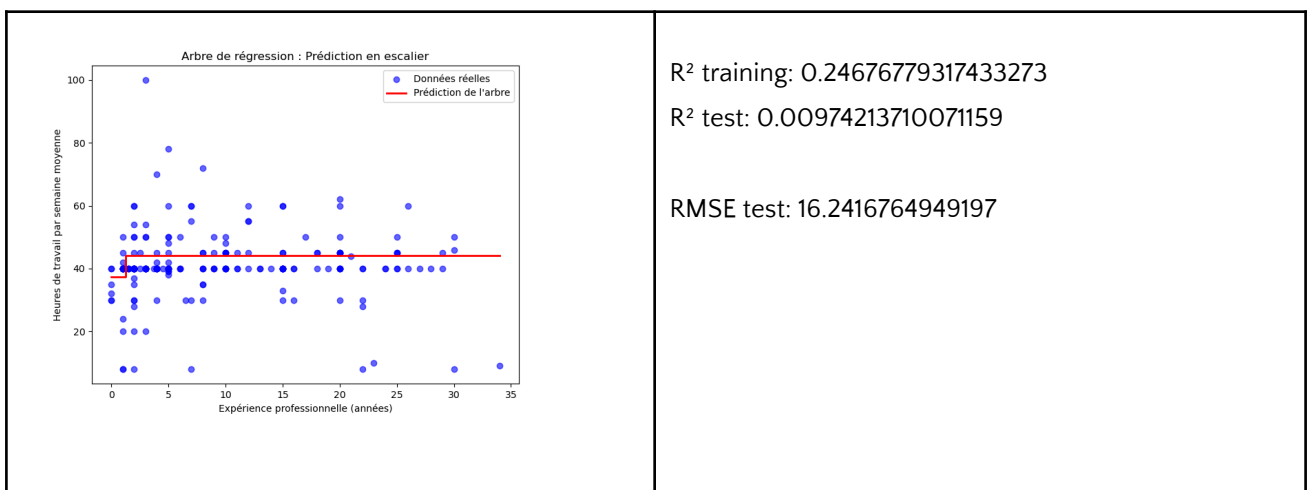


Figure 12: A regression tree

In a graphical representation of a regression tree, the variable Y is examined as a function of a single variable X^i , while all other variables are replaced by constants (their respective means). This yields a step function representation.

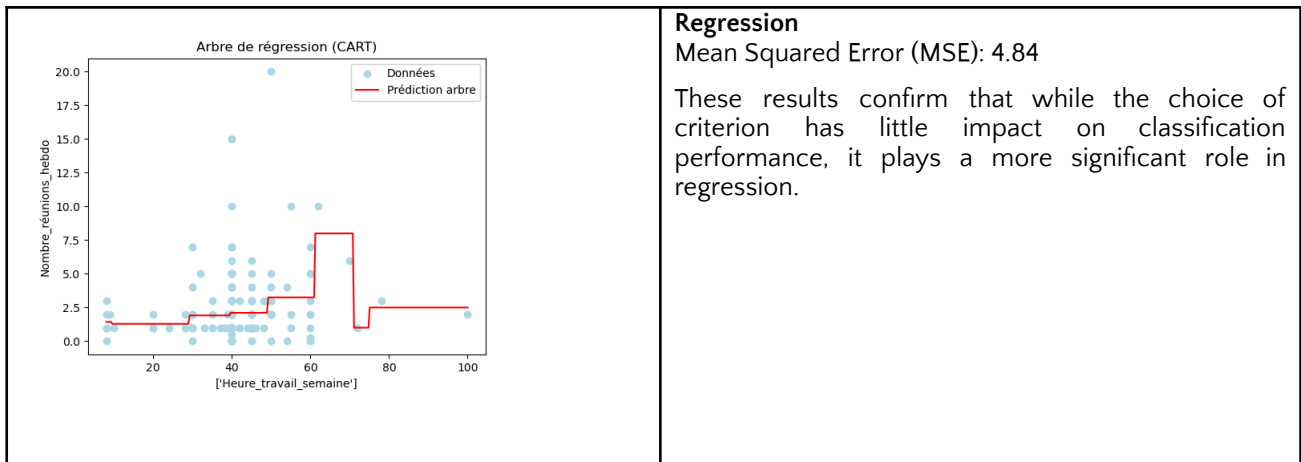


Figure 13: The choice of criterion on regression performance.

3.3 Methodological Perspectives in Machine Learning with trees

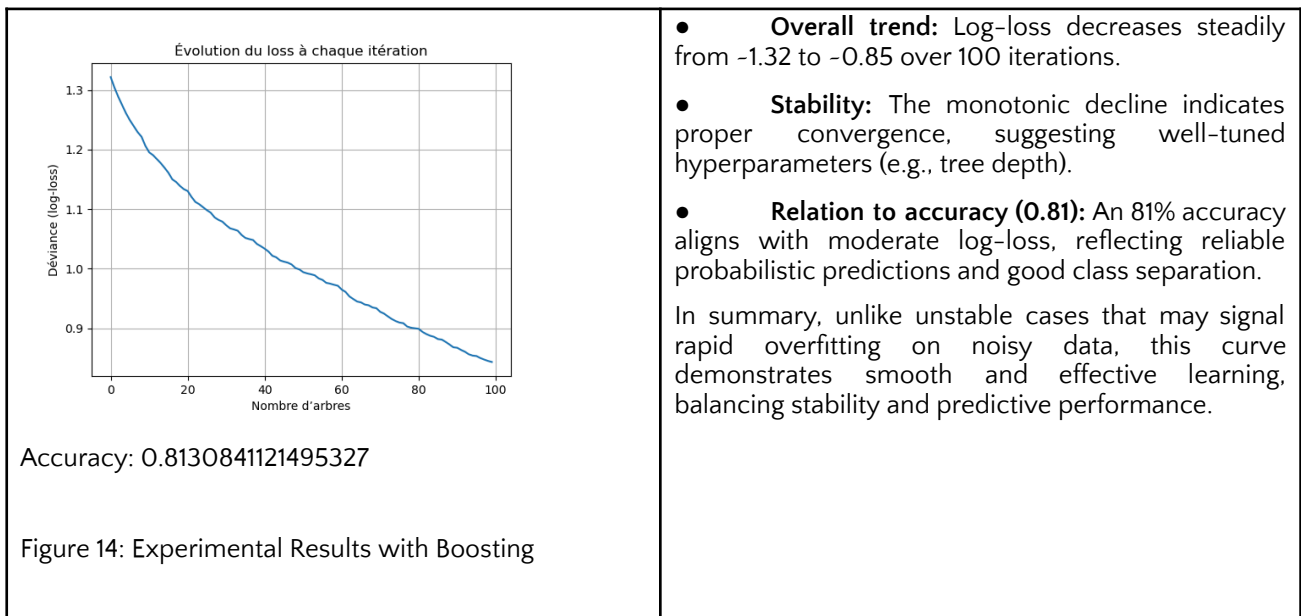


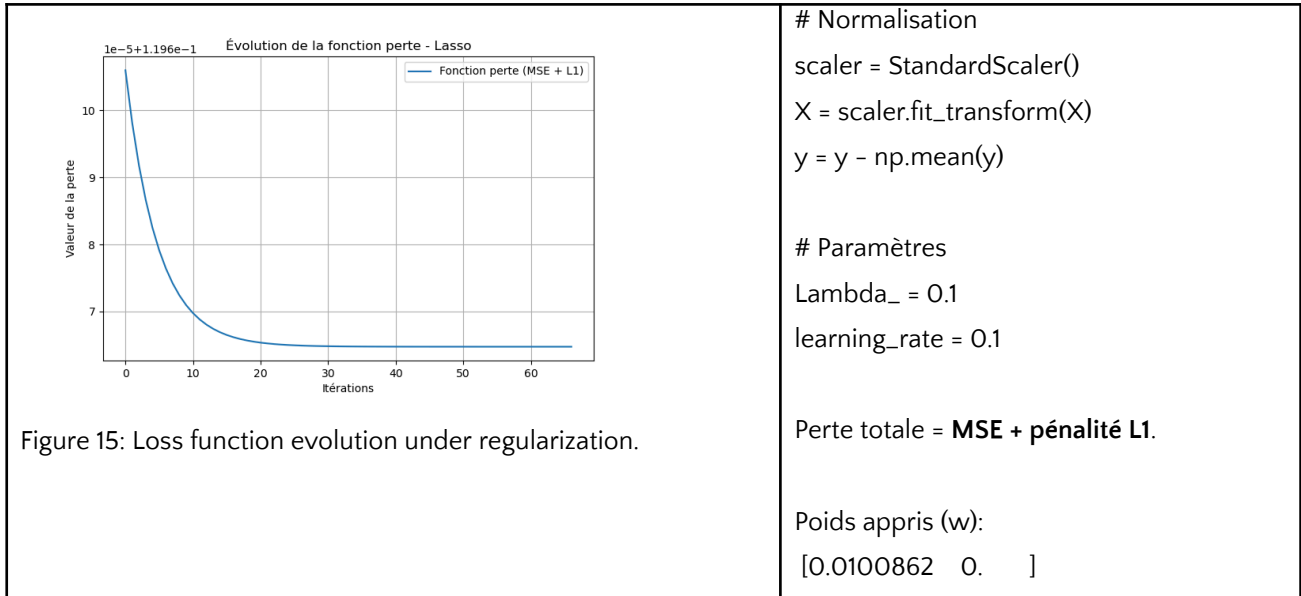
Figure 14: Experimental Results with Boosting

A significant improvement in accuracy was achieved (0.81) using the Gradient Boosting method, which aggregates shallow decision trees with a learning rate of 0.1. The loss function converged steadily.

3.4 Decoupling between variable selection and model specification,

The data **were standardized** (X centred and scaled, y centred), which is important to ensure fair regularization across variables, which is crucial for algorithms sensitive to distances (e.g., k -NN, SVM, linear regression with regularization).

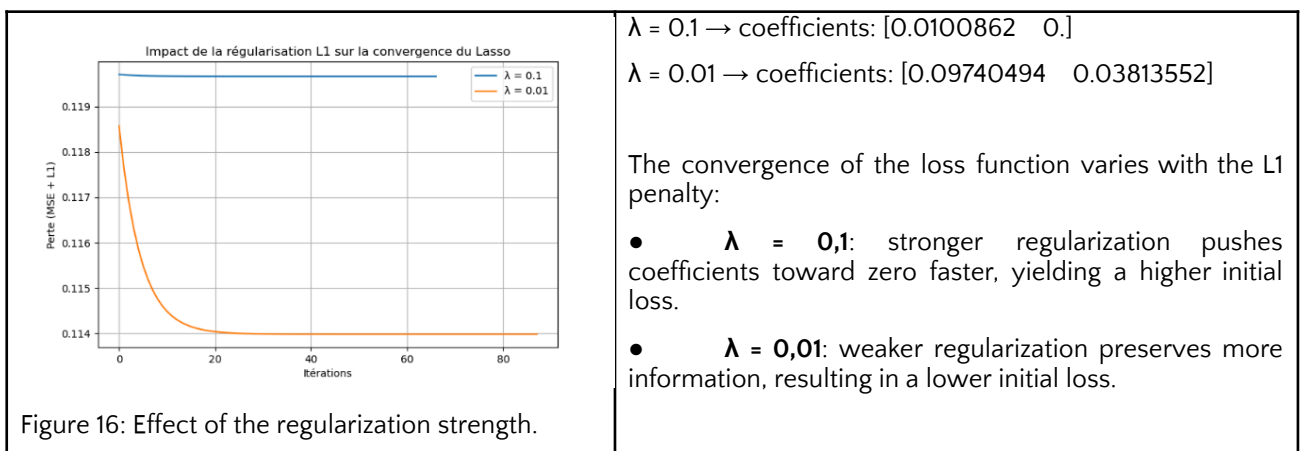
In this study, we applied **frequency encoding** to transform categorical variables into numerical ones. Each category of a qualitative variable was replaced by its relative frequency in the dataset. This approach allowed us to retain the information contained in the categories while making the variables compatible with machine learning algorithms, which require numerical inputs. The loss function, defined as total loss = MSE + L1 penalty, balances predictive accuracy with model sparsity (Tibshirani, 2021)



The first variable ("*13.exp_pro*") has a very small coefficient (≈ 0.01), so its influence on the prediction is nearly negligible. The second variable ("*15.Taille_équipe*") has been completely shrunk to 0.0, which is an expected outcome of Lasso, as it performs variable selection.

Successes include the correct transformation (standardization) and proper application of regularization (variable selection).

Limitations for predictive model quality may arise from overly strong regularization or the selected variables not carrying sufficient informative content.



3.4.1 When More Neurons Don't Mean Better Models: The Importance of Hyperparameter Optimization

One approach to improve model performance is to increase the number of layers and neurons, but experience shows that this is not always the optimal solution.

A multilayer perceptron with 10 hidden neurons was implemented, this hyperparameter defining the model's capacity to capture complex data relationships.

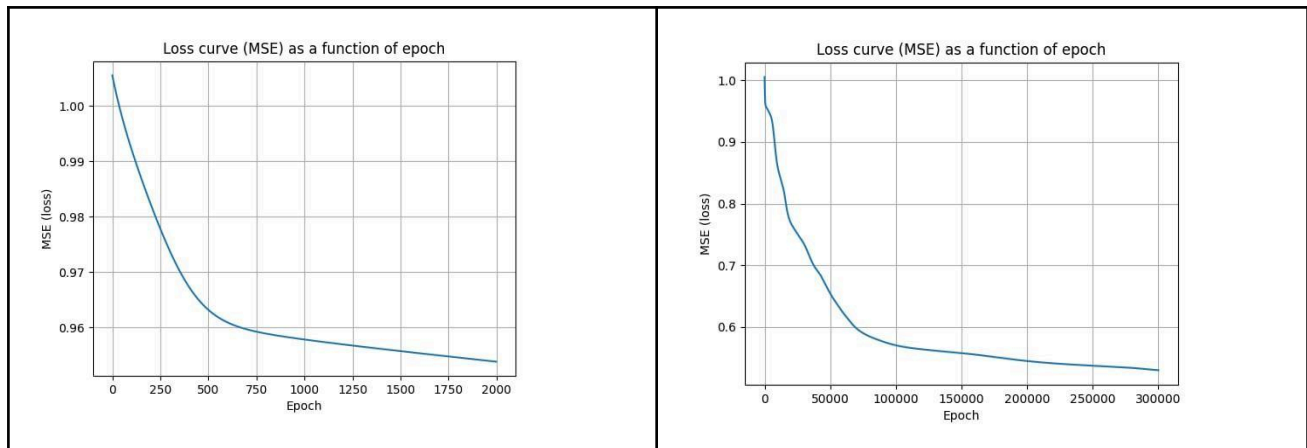


Figure 17: Loss function evolution over epochs

In the first figure, the epoch goes up to 2000 and the MSE is about 0.955. In the second figure, the epoch goes up to 300,000 and the MSE is about 0.53. In the first figure, the epoch goes up to 2000 and the MSE is about 0.95. In the second figure, the epoch goes up to 300,000 and the MSE is about 0.53.

We observe that as the number of epochs increases, the value of the MSE (Mean Squared Error) decreases. This means that the neural network gradually learns to better predict the target variable by adjusting its parameters to reduce the error between the predictions and the actual values. This decreases in MSE over the epochs illustrates the effectiveness of the training process.

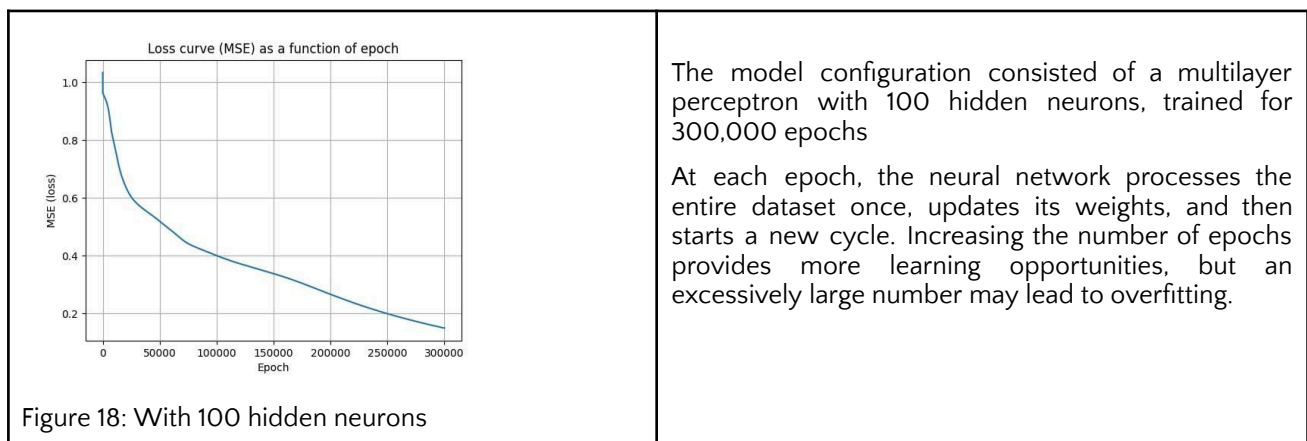


Figure 18: With 100 hidden neurons

The model configuration consisted of a multilayer perceptron with 100 hidden neurons, trained for 300,000 epochs

At each epoch, the neural network processes the entire dataset once, updates its weights, and then starts a new cycle. Increasing the number of epochs provides more learning opportunities, but an excessively large number may lead to overfitting.

4 Discussion, Conclusions, and Perspectives

Regarding model selection, tree-based approaches (decision trees, Random Forests, Gradient Boosting) clearly outperformed gradient descent methods (GD, SGD, mini-batch) and SVMs. Evaluation was based on loss curve progression and test set accuracy. **Key advantages of tree-based models include:**

- Capturing complex nonlinear relationships and variable interactions naturally.
- Robustness to unscaled variables, reducing preprocessing requirements.
- Limited sensitivity to outliers.
- Simultaneous handling of quantitative and qualitative variables.
- Improved generalization on heterogeneous or noisy datasets through ensemble aggregation.

Methods such as L1 regularization (Lasso) and tree-based models (Random Forest, Gradient Boosting) effectively achieve both objectives: identifying informative variables, capturing complex nonlinear relationships, and producing stable models. This decoupling of variable selection and model specification is a major step toward robust, generalizable, and scientifically exploitable predictive models in employability research.

These findings provide a foundation for future studies to validate and extend the adopted approaches. For future work, we plan to explore **regularized tree boosting (Chen & Guestrin, 2016)** and we will investigate the new fitting approaches in a more extensive simulation study like **Random-squared Forests (Kraabel et al., 2020)**.

This study on the employability of young management graduates in Madagascar identified key factors that facilitate labour market integration. Technical, digital, and interpersonal skills emerged as essential, while sector

professionals highlighted the importance of integrating new technologies and AI into curricula. Personal qualities were also emphasized as critical for employability.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Biographies:

Ravelonahina Hasina is a lecturer at INSCAE (Institut National des Sciences Comptables et de l'Administration des Entreprises), teaching Linear Algebra, Optimization, Operations Research, and Data Analysis. He also teaches Mathematics at Lycée Français de Tananarive Ambatobe. He is pursuing a Ph.D. at Doctoral School of Science and Engineering and Technic Innovation (STII), (Laboratory of Cognitive Science and Application. University of Antananarivo, Madagascar), focusing on artificial intelligence and high-dimensional optimization algorithms. His research experience includes conducting social surveys on topics such as COVID-19 impacts, cost of living, and governance perceptions among students in Antananarivo.

ROBINSON Matio: Doctor HDR (Habilitation to Supervise Research) in STII, University of Antananarivo.

ANDRIAMANOHISOA Hery Zo: Professor in STII, University of Antananarivo.

References:

- Bottou, L. (1991). Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nimes*, 91(8), 12.
- Breiman, L. (2001a). Random Forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3), 199-231.
- Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (2017). *Classification and Regression Trees*. Chapman and Hall/CRC. <https://doi.org/10.1201/9781315139470>
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. <https://doi.org/10.1145/2939672.2939785>
- Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4), 367-378.
- Krabel, T. M., Tran, T. N. T., Groll, A., Horn, D., & Jentsch, C. (2020). *Random boosting and random² forests—A random tree depth injection approach* (No. arXiv:2009.06078). arXiv. <https://doi.org/10.48550/arXiv.2009.06078>
- Krzywinski, M., & Altman, N. (2017). Classification and regression trees. *Nature methods*, 14(8), 757-758. <https://doi.org/10.1038/nmeth.4370>
- Mayr, A., Binder, H., Gefeller, O., & Schmid, M. (2014). The Evolution of Boosting Algorithms: From Machine Learning to Statistical Modelling. *Methods of Information in Medicine*, 53(06), 419-427. <https://doi.org/10.3414/ME13-01-0122>
- Moumen, A., Bouchama, E. H., & EL IDIRISSI, Y. E. B. (2020). Data mining techniques for employability: Systematic literature review. *2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, 1-5. <https://ieeexplore.ieee.org/abstract/document/9314555/>
- Tibshirani, R. J. (2021). Equivalences between sparse models and neural networks. *Working Notes*. URL <https://www.stat.cmu.edu/ryantibs/papers/sparsitynn.pdf>