Heliyon 12 (2026) e45629

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

STEM employment in Spain: An explainable machine learning analysis from a national survey

Luis Ángel Calvo Pascual [a,b,*], Ana Zapatero González [a]

[a] Department of Quantitative Methods, ICADE, Universidad Pontificia de Comillas, C/Alberto Aguilera, 23, Madrid, Spain
[b] Institute for Research in Technology (IIT), Madrid, Spain

ARTICLE INFO

Keywords:
STEM occupation
Bayesian optimization
Association rule mining

ABSTRACT

This study examines the variables associated with STEM (science, technology, engineering, and mathematics) employment using the 2019 survey on the labour market insertion of Spanish university graduates. The outcome distinguishes graduates employed in a STEM occupation about five years after graduation from all other graduates, whether employed in non-STEM occupations or not employed. Using explainable machine learning methods (feature selection, a Bayesian-optimized decision tree and association rule mining), we find that, among characteristics fixed at graduation, holding a STEM degree is the strongest predictor of STEM employment, followed by gender: a tree based only on these characteristics reduces to a single split on the STEM degree and correctly classifies 77% of the graduates employed in STEM occupations and 84% of the others. Advanced ICT skills and graduates' retrospective assessment that their degree helped them to find a job, both reported in 2019, are also associated with STEM employment. Gender differences are consistent with the lower participation of women in STEM degrees and their lower rate of STEM employment among STEM-degree holders, and the main results hold when the analysis is restricted to employed graduates.

1. Introduction

Extensive research has examined the labour-market outcomes of university graduates, including differences between STEM (science, technology, engineering, and mathematics) and non-STEM employment in the United States [1], the role of the field of study in the employability of young graduates in Europe [2], and cross-national differences in graduate recruitment and graduate labour markets [3,4]. Other studies have identified employability profiles among graduates of a Spanish university, including one based on information technology expertise [5], and have reported a positive association between STEM-educated workers and output growth in Europe, where the share of STEM graduates among tertiary graduates has declined over the past two decades [6]. This evidence motivates a closer examination of the factors associated with entering STEM occupations. While various studies have explored why young people choose to study STEM [7], little research has employed machine learning methods to analyze the characteristics of Spanish university graduates working in STEM positions. Such an approach makes it possible to identify the variables most strongly associated with graduates' transition into STEM employment.

This study uses explainable machine learning models, including feature selection (mutual information and four complementary importance measures), Bayesian-optimized decision trees and association rules, to identify the variables most strongly associated with

* Corresponding author. Department of Quantitative Methods, ICADE, Universidad Pontificia de Comillas, C/Alberto Aguilera, 23, Madrid, Spain.
E-mail address: lacalvo@comillas.edu (L.Á. Calvo Pascual).

https://doi.org/10.1016/j.heliyon.2026.e45629
Received 24 February 2026; Received in revised form 13 October 2026; Accepted 21 October 2026
2405-8440/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

STEM employment among Spanish university graduates. This approach avoids the black-box nature of many machine learning models while retaining reasonable predictive performance and capturing non-linear relationships and interactions. Our findings highlight the central role of holding a STEM degree, together with self-reported advanced ICT skills, gender differences and graduates' retrospective assessment of whether their degree helped them to find a job. Sensitivity analyses restricted to characteristics fixed at graduation and to employed graduates confirm the predominance of the STEM degree, followed by gender. The analysis is predictive and associative; it does not identify causal effects.

This article is organized as follows. Section 2 provides a literature review. In Section 3, we present the data and methodology, describing the source, the sample characteristics, and the explainable machine learning techniques applied. Next, in Section 4, we discuss the results, focusing on both the predictive performance and interpretability of the models, reporting sensitivity analyses and comparing the results with the existing literature. Finally, in Section 5, we conclude by summarizing the main contributions of this study and offering policy recommendations.

2. Literature review

The transition from higher education to employment has received considerable attention in the European literature. Although the expansion of tertiary education has increased the supply of highly qualified workers, the growth of graduate-level jobs has been slower and uneven across countries [4]. Previous research has documented mismatches between educational qualifications and labour-market requirements, including overeducation and skill mismatches [8,9] and discrepancies between graduates' fields of study and their occupations [10]. These findings suggest that obtaining a university degree and securing employment aligned with one's academic training are distinct labour-market outcomes. The latter is particularly relevant when analysing the transition from STEM education to STEM occupations.

STEM education occupies a prominent position in discussions about technological development, productivity and the demand for specialised skills. At the European level, Bacovic et al. [6] report a strong descriptive association between science and technical education, the structure of output, income per capita and productivity growth. However, the relationship between education and employment should also be examined at the individual level. López López et al. [7], in a systematic mapping of the literature, classify the factors associated with STEM career choices into personal, social and environmental dimensions. Their review highlights the prominence of psychological factors in the literature on STEM degree choice and the particular attention given to career prospects in studies of Computer Science. Nevertheless, the decision to study a STEM discipline does not necessarily imply subsequent employment in a STEM occupation, making the transition between education and occupational outcomes a separate question of interest.

In Spain, the relationship between university education and labour-market demand has been examined from several perspectives. Lacuesta et al. [11] analyze mismatches between the demand for and supply of places in Spanish public in-person universities and find that the degrees with the greatest access difficulty are those whose graduates earn the highest salaries; their review of the international literature indicates that degree choices depend on expected earnings, perceived skills and personal preferences, with subjective expectations playing a dominant role. Official data from the 2019 Survey on the Labour Market Insertion of University Graduates (EILU) also show substantial differences in employment outcomes by field of study: for graduates of the 2013–2014 academic year, the employment rate in 2019 reached 92.4% in engineering and architecture but 83.9% in the sciences, compared with 86.1% for all graduates [12]. These figures refer to employment by field of study rather than to employment in STEM occupations, and their heterogeneity across STEM-related fields illustrates the importance of examining educational specialisation and subsequent occupational outcomes separately.

Gender differences represent another important dimension of STEM educational and occupational trajectories. Previous studies have documented women's underrepresentation in STEM programmes in Spain and the European Union [13], their underrepresentation in ICT education and digital occupations in Europe [14], and gender gaps in salary and employment status among STEM graduates in the United States [15]. In Spain, in the 2019–2020 academic year, 13 of every 100 women enrolled in bachelor's programmes (ISCED level 6) had chosen STEM fields, compared with 16 of every 100 in the European Union [13]. Valdés and Solga [16] provide direct evidence of a gender gap in the transition from STEM education to STEM employment in Spain: female STEM graduates are less likely than male graduates to work in STEM occupations, both in their first job and four to five years after graduation, and the gap is larger among those who graduated during the economic crisis and among those from low socioeconomic backgrounds. Complementary research has examined earlier stages of STEM pathways; for example, González-Pérez et al. [17] find that exposure to female STEM role models is associated with improvements in girls' mathematics-related expectations and STEM career aspirations. Together, these studies distinguish between gender differences in educational choices and inequalities in subsequent occupational trajectories.

Beyond academic specialisation and demographic characteristics, the alignment between higher education and labour-market needs has also been highlighted as relevant to students' post-college outcomes [18]. Digital competencies may be particularly relevant for STEM occupations, which include ICT professionals and technicians. However, educational qualifications, self-reported competencies and retrospective evaluations of the usefulness of a degree represent conceptually different types of information. In graduate surveys conducted several years after graduation, reported competencies and evaluations of educational usefulness may partly reflect experiences accumulated after entering the labour market. Their associations with employment outcomes should therefore not be interpreted as evidence of temporal precedence or causal effects.

Taken together, the literature provides substantial evidence on STEM educational choices, graduate labour-market mismatches and gender inequalities in STEM careers. The present study complements these contributions by examining the relative importance and joint associations of educational, demographic and skill-related characteristics with STEM employment among Spanish university

graduates. Using a large national survey, we combine feature-selection methods, Bayesian-optimized decision trees and association rule mining to obtain interpretable patterns of STEM employment. Furthermore, we distinguish between characteristics available at graduation and information reported after labour-market entry, and we assess the stability of the results when the predictive analysis is restricted to antecedent variables and to employed graduates. This approach provides a complementary predictive and associative

Table 1
Descriptive summary of variables: frequencies, percentages, and missing values.

Variable	Role in the analysis	Main distribution	Missing cells
STEM occupation	Target variable	Employed in a STEM occupation: 6217 (19.64%); not employed in a STEM occupation: 25,434 (80.36%), of which employed in a non-STEM occupation: 20,907 (66.05%) and not employed: 4527 (14.30%)	0 (0.00%)
STEM degree	Main educational predictor	STEM degree: 8918 (28.18%); non-STEM degree: 22,733 (71.82%)	0 (0.00%)
First STEM employment	Early labour-market outcome	First STEM employment: 5907 (18.66%); non-STEM first employment: 25,744 (81.34%)	0 (0.00%)
Gender	Demographic variable	Women: 18,056 (57.05%); men: 13,595 (42.95%)	0 (0.00%)
Age	Demographic variable	Under 30: 15,707 (49.63%); 30–34: 8825 (27.88%); 35 or over: 7119 (22.49%)	0 (0.00%)
Country	Demographic variable	Spain: 30,758 (97.18%); outside the EU: 616 (1.95%); EU: 277 (0.88%)	0 (0.00%)
Qualification	Field of study	Most frequent categories: Administration and business 1988 (6.28%); Primary education 1907 (6.03%); Law 1720 (5.43%); Early childhood education 1365 (4.31%); Nursing 1235 (3.90%). Full distribution in S1.	0 (0.00%)
Type of university	Academic background	Public in-person: 25,898 (81.82%); private in-person: 3793 (11.98%); public distance: 1080 (3.41%); private distance: 880 (2.78%)	0 (0.00%)
Scholarship	Academic background	No: 16,866 (53.29%); yes: 14,759 (46.63%); don't know/no answer: 26 (0.08%)	0 (0.00%)
Studies abroad	Academic background	No: 26,340 (83.22%); yes: 5291 (16.72%); don't know/no answer: 20 (0.06%)	0 (0.00%)
Study motivation (reason for choosing the degree)	Academic motivation	Future of work: 23,673 (74.79%); expand knowledge: 5090 (16.08%); other reasons: 2773 (8.76%); don't know/no answer: 115 (0.36%)	0 (0.00%)
Degree satisfaction (would study the same degree again)	Retrospective evaluation of the degree	Asked only of graduates who would study at university again. Yes: 19,469 (71.70% of valid responses); no: 7571 (27.88%); don't know/no answer: 114 (0.42%)	4497 (14.21%)
Degree usefulness (degree helped to find a job)	Perceived labour-market usefulness	Yes: 22,003 (71.04% of valid responses); no: 8709 (28.12%); don't know/no answer: 260 (0.84%)	679 (2.15%)
ICT skills	Skill-related variable	Intermediate: 21,240 (67.11%); advanced: 6596 (20.84%); basic: 3766 (11.90%); don't know/no answer: 49 (0.15%)	0 (0.00%)
Language proficiency	Skill-related variable	High: 14,412 (48.81% of valid responses); intermediate: 12,734 (43.13%); low: 2142 (7.26%); don't know/no answer: 236 (0.80%)	2127 (6.72%)
Language certification	Skill-related variable	Yes: 17,084 (57.86% of valid responses); no: 12,350 (41.83%); don't know/no answer: 90 (0.30%)	2127 (6.72%)
Current employment	Labour-market status	Working: 27,124 (85.70%); unemployed: 2438 (7.70%); inactive: 2089 (6.60%)	0 (0.00%)
Professional status	Current employment conditions	Permanent salaried employee: 15,558 (57.36% of valid responses); temporary salaried employee: 7572 (27.92%); self-employed without employees: 1832 (6.75%); trainee/intern: 1508 (5.56%)	4527 (14.30%)
Work schedule	Current employment conditions	Full time: 23,363 (86.13% of valid responses); part time: 3761 (13.87%)	4527 (14.30%)
Current occupation	Labour-market outcome	Most frequent categories: Teaching professionals 4216 (15.54% of valid responses); health professionals 3375 (12.44%); scientists and engineers 2907 (10.72%); financial specialists 1733 (6.39%); engineering technicians 1051 (3.87%). Full distribution in S1.	4527 (14.30%)
Current salary	Labour-market outcome	Most frequent categories: €1000–1499: 8415 (31.02% of valid responses); €1500–1999: 7887 (29.08%); €2000–2499: 3442 (12.69%); €700–999: 2524 (9.31%)	4527 (14.30%)
First occupation	First labour-market outcome	Most frequent categories: health professionals 3157 (10.19% of valid responses); teaching professionals 2812 (9.08%); scientists and engineers 2743 (8.86%); support services 1704 (5.50%); shop assistants 1693 (5.47%). Full distribution in S1.	679 (2.15%)
Status of first employment	First labour-market outcome	Temporary salaried employee: 12,376 (39.96% of valid responses); permanent salaried employee: 10,321 (33.32%); trainee: 5927 (19.14%); self-employed: 1582 (5.11%)	679 (2.15%)
First salary	First labour-market outcome	Less than €700: 9627 (31.08% of valid responses); €1000–1499: 7912 (25.55%); €700–999: 7100 (22.92%); €1500–1999: 3593 (11.60%)	679 (2.15%)
Theoretical factors (most appropriate field of study for the current job)	Educational/labour-market alignment	No particular area: 6870 (39.81% of valid responses); totally different area: 4275 (24.77%); own or related area: 2971 (17.22%); exclusively its area: 2809 (16.28%)	14,395 (45.48%)
Practical factors	Educational/labour-market alignment	A lot: 10,164 (41.25% of valid responses); quite: 6976 (28.31%); some: 3034 (12.31%); nothing: 2155 (8.75%); a little: 1967 (7.98%)	7013 (22.16%)
ICT factors	Educational/labour-market alignment	Quite: 7084 (28.75% of valid responses); some: 5424 (22.01%); a lot: 5299 (21.51%); a little: 3391 (13.76%); nothing: 3066 (12.44%)	7013 (22.16%)

perspective without attributing causal effects to the identified relationships.

3. Data and methodology

3.1. Data

Our dataset is derived from the 2019 Survey on the Labour Market Insertion of University Graduates (EILU 2019) of the Spanish National Statistics Institute [12], which interviewed the graduates of the 2013–2014 academic year about five years after graduation, comprising a sample of 31,651 graduates and 40 variables, including demographic variables (e.g., gender, age, country, marital status, cohabitation with a partner), academic variables (STEM degree, type of university (private or public), ICT skills, foreign language level, degree satisfaction, measured by whether respondents would study the same degree again (asked only of those who would study at university again), and degree usefulness, measured by whether respondents believe that their degree has helped them to find a job), and economic variables (STEM occupation, current salary, first salary, social security contributions). The prediction target is the binary variable STEM occupation, which distinguishes graduates employed in a STEM occupation at the time of the interview (class 1) from all other graduates (class 0), that is, graduates employed in a non-STEM occupation and graduates who were not employed (unemployed or inactive). In the sample, 6217 graduates (19.64%) were employed in STEM occupations, 20,907 (66.05%) in non-STEM occupations and 4527 (14.30%) were not employed. The models distinguish being employed in a STEM occupation from not being employed in one; a sensitivity analysis restricted to employed graduates, which compares STEM with non-STEM employment only, is reported in Section 4.1.4. The analytical file retains 32 predictors, which are listed with their coding in Supplementary Material S6. The EILU produces population estimates with elevation factors calibrated to population totals [12]. The analyses reported here do not use these sampling weights; all proportions, models and performance metrics therefore refer to the unweighted analytical sample rather than to the population of 2013–2014 graduates.

To make the main characteristics of the sample transparent within the manuscript, Table 1 provides a reduced descriptive summary of the key variables used in the analysis, including their main categories, frequencies, percentages, and missing values. The complete descriptive summary is provided in Supplementary Material S1_descriptive.docx.

STEM degrees were identified from the Spanish classification of fields of education (CNED-2014), which is aligned with ISCED-F 2013 [19], and comprise programmes in the natural sciences, mathematics and statistics, information and communication technologies, and engineering, manufacturing and construction, together with a small number of engineering-oriented programmes classified in other broad fields (e.g., agricultural, forestry and biomedical engineering). STEM occupations were identified from the Spanish classification of occupations (CNO-11) and comprise science, engineering and ICT professionals and technicians, together with related managerial, supervisory and mathematical support occupations. Medicine and other health degrees and occupations were excluded, consistent with definitions that classify health care outside STEM [20]; two exceptions with a strong engineering or physical-science component were retained: biomedical engineering, which these definitions also include in STEM [20], and optics and optometry. The inclusion of health-related fields nevertheless varies across definitions [20,21]. The complete classification of STEM degrees and occupations is provided in Supplementary Material S2_STEM_definition.docx.

3.2. Methodology

We use three different machine learning approaches (feature selection, a Bayesian-optimized decision tree, and association rule mining) to balance interpretability and predictive performance. Feature selection ranks the variables according to their association with the outcome, the Bayesian-optimized decision tree provides a concise predictive model, and association rule mining complements

Table 2
Summary of the objective, inputs, outputs, and role of each analytical method.

Method	Objective	Input data	Output	Role in the study
Feature selection	Identify the variables most strongly associated with the binary target variable STEM occupation.	Preprocessed dataset including demographic, academic and labour-market variables; target variable: STEM occupation.	Normalized feature-importance scores and a ranked list of predictors across mutual information, random forest, logistic regression, chi-square and XGBoost.	Provides an exploratory, descriptive screening of associations (including variables measured after graduation); not a selection of predictors available at graduation.
Bayesian-optimized decision tree	Build an interpretable predictive model for STEM occupations while retaining predictive performance.	Nine education-, skill- and gender-related predictors (labour-market outcome variables excluded); 70/30 stratified train/test split; binary target variable: STEM occupation.	Predicted class (employed in a STEM occupation or not), optimized tree structure, accuracy, precision, recall, F1-score and confusion matrix.	Serves as the main explainable predictive model and identifies the hierarchical decision rules associated with STEM employment.
Association rule mining	Detect combinations of educational, demographic and skill-related factors associated with STEM labour-market outcomes.	One-hot encoded categorical variables; antecedents restricted to educational, demographic and skill-related factors; consequent: STEM occupation = yes (joint labour-market outcomes in S5).	Association rules described by support, confidence and lift and ranked by lift, plus a variable-exclusion analysis.	Complements the decision tree by revealing multi-variable patterns and interpretable antecedent-consequent relationships.

both by identifying combinations of variable levels that co-occur with STEM employment. In the following sections, we detail the methodology of each of these methods; Table 2 summarizes the objective, input data, output and role of each method.

Prediction time point and timing of the predictors. The EILU collects all information retrospectively in a single interview, about five years after graduation. We define the intended prediction time point as the end of university studies (2014) and the outcome as the occupation held at the time of the interview (2019). Table 3 classifies the 32 predictors according to the period to which they refer: (i) characteristics determined before or at graduation (gender, age group, STEM degree, type of university, scholarship, studies abroad and reason for choosing the degree), together with disability status and country, which were recorded at the interview but are unlikely to have changed since graduation for most respondents (age group is a deterministic function of the year of birth); (ii) skills and degree evaluations that do not refer to a specific job but were self-reported in 2019 (ICT skills, foreign language, language proficiency and certification, and degree satisfaction, whose values may have changed after graduation and therefore cannot be assumed to have been available at the prediction time point; (iii) the post-graduation labour-market history (first STEM employment, status of the first job, use of the knowledge acquired in the first job, work experience, job rejection, social security contributions and degree usefulness, i.e. whether the degree helped to find a job); and (iv) variables measured for the current job or concurrently with the outcome (current employment, professional status, work schedule, multiple employment, current salary, the field of study most appropriate for the current job, the influence of practical skills, languages and ICT training on obtaining the current job [practical, language and ICT factors], marital status and cohabitation). Only group (i) can be regarded as available at the prediction time point. Variables in groups (iii) and (iv) refer to the period after graduation, and some of them are partly determined by the outcome; the questions on the current job, for instance, are asked only of employed respondents, so that their imputed values also partly identify graduates who are not employed. The labour-market outcomes in these groups (e.g., first STEM employment, current employment, salary, work schedule and professional status) are therefore used only in the descriptive screening of Fig. 1 and in the benchmark set D (Section 4.1.4); degree usefulness and the four current-job factors are predictors of the main decision tree (set C) and antecedents of the association rules, and their associations are interpreted accordingly. Sensitivity analyses restricted to group (i) (set A) and to groups (i) and (ii) (set B) are reported in Section 4.1.4.

Data preparation and encoding. In the analytical file, all variables are stored as integer codes: binary (yes/no) variables are coded 0/1 (1 = yes; for gender, 1 = male and 0 = female); multi-category variables are label-encoded, with codes assigned in the

Table 3
Timing of the predictors relative to the prediction time point (graduation) and mean normalized importance by predictor set. Set A: characteristics determined before or at graduation (disability and country were recorded at the 2019 interview); set B: set A plus skills and degree evaluations reported in 2019 that do not refer to a specific job and may have changed after graduation; set C: predictors of the main decision tree; set D: all predictors. Importances (explicit coding; Section 3.2) are computed on the training data within cross-validation folds and sum to one within each set; – = variable not included.

Variable	Timing	Set A	Set B	Set C	Set D
STEM Degree	(i) Before/at graduation	0.836	0.629	0.498	0.136
Gender	(i) Before/at graduation	0.077	0.052	0.043	0.013
Age	(i) Before/at graduation	0.021	0.024	–	0.006
Study Motivation	(i) Before/at graduation	0.011	0.015	–	0.006
Type of University	(i) Before/at graduation	0.016	0.014	–	0.005
Scholarship	(i) Before/at graduation	0.013	0.016	–	0.004
Studies Abroad	(i) Before/at graduation	0.013	0.012	–	0.003
Country	(i) Before/at graduation	0.006	0.006	–	0.002
Disability	(i) Before/at graduation	0.006	0.004	–	0.001
ICT Skills	(ii) Skills/evaluations, reported 2019	–	0.116	0.077	0.025
Language	(ii) Skills/evaluations, reported 2019	–	0.037	–	0.013
Language Certification	(ii) Skills/evaluations, reported 2019	–	0.027	–	0.010
Degree Satisfaction	(ii) Skills/evaluations, reported 2019	–	0.022	0.017	0.007
Language Proficiency	(ii) Skills/evaluations, reported 2019	–	0.025	–	0.006
First STEM Employment	(iii) Post-graduation history	–	–	–	0.355
Status of First Employment	(iii) Post-graduation history	–	–	–	0.024
Social Security Contributions (Last 5 Years)	(iii) Post-graduation history	–	–	–	0.012
Work Experience	(iii) Post-graduation history	–	–	–	0.012
Degree Usefulness	(iii) Post-graduation history	–	–	0.042	0.010
Knowledge at First Employment	(iii) Post-graduation history	–	–	–	0.010
Job Rejection	(iii) Post-graduation history	–	–	–	0.006
Current Employment	(iv) Current job/concurrent	–	–	–	0.068
Professional Status	(iv) Current job/concurrent	–	–	–	0.060
Work Schedule	(iv) Current job/concurrent	–	–	–	0.045
ICT Factors	(iv) Current job/concurrent	–	–	0.132	0.041
Current Salary	(iv) Current job/concurrent	–	–	–	0.026
Language Factors	(iv) Current job/concurrent	–	–	0.076	0.025
Multiple Employment	(iv) Current job/concurrent	–	–	–	0.021
Practical Factors	(iv) Current job/concurrent	–	–	0.066	0.021
Theoretical Factors	(iv) Current job/concurrent	–	–	0.051	0.020
Marital Status	(iv) Current job/concurrent	–	–	–	0.005
Cohabitation with Partner	(iv) Current job/concurrent	–	–	–	0.004

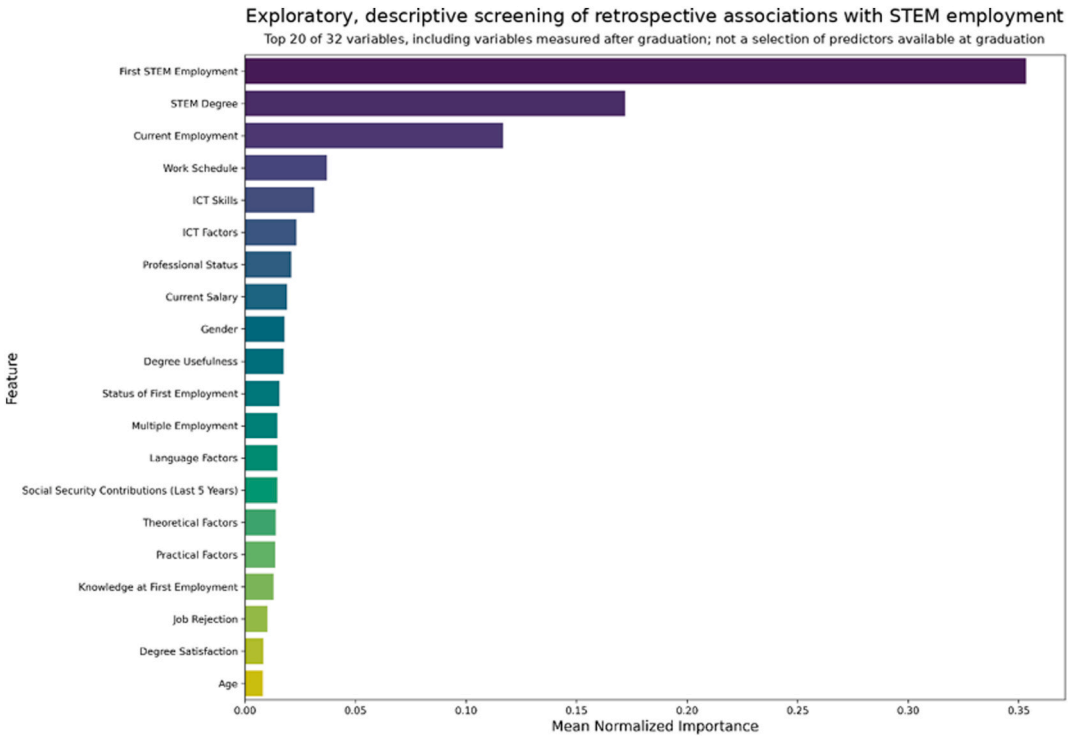


Fig. 1. Exploratory, descriptive screening of retrospective associations with STEM employment; it is not a selection of predictors available at graduation. Bars show the mean normalized importance of the 20 most important of all 32 variables, averaged over mutual information, random forest feature importance, logistic regression coefficients, chi-square test and XGBoost feature importance (the importances of each method are normalized to sum to one before averaging). The highest-ranked variables include labour-market variables measured after graduation or concurrently with the outcome (e.g., first STEM employment, current employment, work schedule, professional status and current salary), which describe the employment trajectory itself. See Table 3 for the timing of each variable and Section 4.1.4 for the antecedent-only analyses on which the main conclusions rest.

alphabetical order of the original Spanish labels (e.g., ICT skills: 0 = advanced, 1 = basic, 2 = intermediate); current salary is coded in seven ordered bands; age is coded by the values 25, 32 and 50 for its three groups (under 30, 30–34, 35 or over); and social security contributions are numeric. 'Don't know/no answer' responses were treated as missing. The full codebook is provided in Supplementary Material S6. In the feature-selection and decision-tree analyses, these codes entered the models as numeric values (no one-hot encoding), and missing values were imputed with the variable mean. Because of this imputation, the thresholds of the decision trees can take non-integer values: scikit-learn places each threshold midway between adjacent observed values, so that in Fig. 2 the split Degree Usefulness ≤ 0.858 separates the 'yes' answers (code 1) from the 'no' answers (code 0) and the missing values (imputed with the mean of the valid responses, 0.716), and the split ICT Skills ≤ 0.5 isolates advanced ICT skills (code 0). No scaling is required for tree-based models and mutual information; for logistic regression, XGBoost and the chi-square test, variables were standardized (z-scores), and for the chi-square test the standardized variables were further discretized into ten equal-width bins to obtain non-negative values. For association rule mining, every category level was converted into a separate binary item (one-hot encoding, e.g. 'ICT skills = advanced'), and missing or 'don't know' answers did not generate items.

Order of operations and leakage control. In the main analysis, imputation and standardization were applied to the full dataset before the 70/30 train/test split, and the feature screening (Fig. 1) was computed on the full dataset; the decision tree was then tuned with cross-validation on the training set and evaluated once on the test set. To limit information leakage, we re-estimated all predictive analyses with training-only preprocessing, that is, with a pipeline in which the train/test split is performed first; imputation, encoding, scaling and discretization are fitted on the training data only (and, within cross-validation, on each training fold); feature rankings are computed on the training folds only; Bayesian hyperparameter optimization uses cross-validation within the training set; and the test set is used only once, for the final evaluation. This re-estimation does not, however, repeat the choice of the nine predictors of the main tree (set C), which was made after inspecting the feature screening computed on the full dataset; set C is therefore a predefined selection. The additional predictor sets A and B are defined by temporal criteria (Table 3) rather than by the data and thus provide checks that do not depend on this selection step. Besides a replication with the original coding, the re-estimation was repeated with an explicit coding, in which nominal variables are one-hot encoded, ordered variables are recoded to their natural order and missing-value indicators are added. In this version, feature rankings were computed with three repetitions of stratified fivefold cross-validation on the training set; the importances of dummy and missing-indicator columns were summed back to the original variable (for logistic regression, the importance of a multi-column variable is the standard deviation of its contribution to the linear predictor, which equals

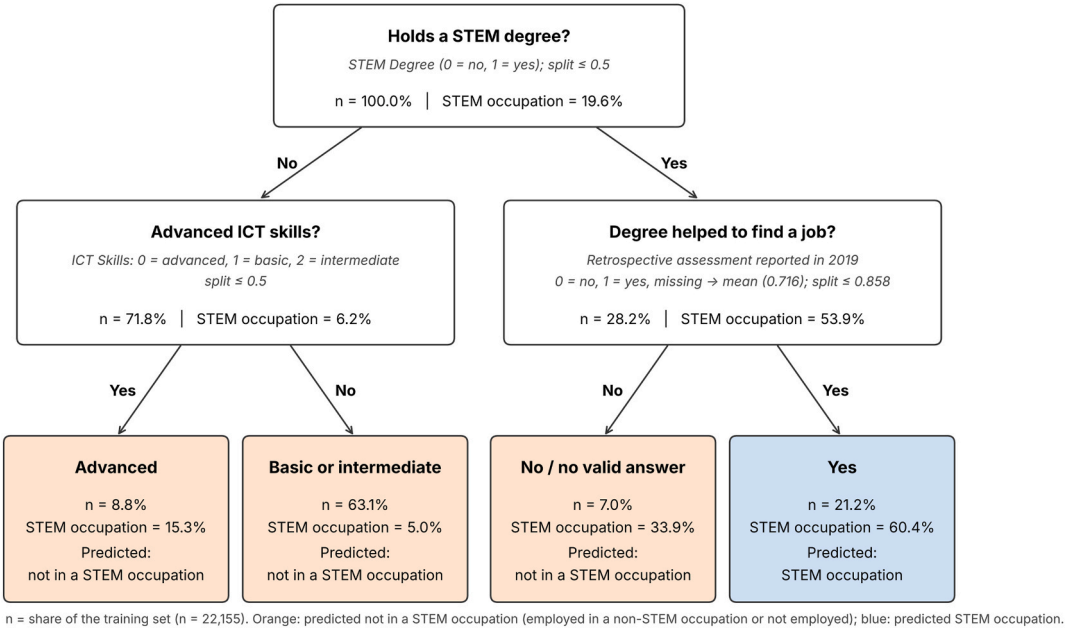


Fig. 2. Pruned decision tree for STEM occupation (maximum depth three; after cost-complexity pruning, depth two and four leaves), based on STEM degree (root), ICT skills (branch without a STEM degree) and degree usefulness, a retrospective assessment reported in 2019 (branch with a STEM degree). Each node shows the share of the training sample (n) and the proportion employed in a STEM occupation; the reference class (not in a STEM occupation) includes graduates employed in non-STEM occupations and graduates who were not employed. The coding of the variables and the meaning of the thresholds are explained in Section 3.2.

the absolute coefficient for a single standardized variable); mutual information used the estimator for discrete variables; gradient-boosting importances were obtained with the scikit-learn implementation using the same hyperparameters as the XGBoost model; and the Bayesian optimization used a Gaussian-process surrogate with expected improvement (10 random and 10 guided evaluations). Results are reported in Section 4.1.4.

3.2.1. Feature selection

One of the goals of this paper is to identify the variables most strongly associated with the binary variable STEM occupation among graduates in Spain. Feature selection is a widely used strategy in machine learning to improve interpretability, reduce dimensionality, and identify the variables most relevant to the target outcome [22]. To achieve this, we employed five distinct feature selection methods: mutual information, random forest feature importance, logistic regression coefficients, chi-square test and *XGBoost* feature importance, using the Python libraries *scikit-learn* and *XGBoost*. These particular techniques were chosen for their complementary strengths and ability to capture different aspects of feature relevance. For example, mutual information captures non-linear dependencies. Random forest feature importance leverages ensemble learning to assess the importance of features based on their contribution to reducing impurity across multiple decision trees, effectively handling feature interactions and correlations [23]. On the other hand, logistic regression coefficients capture linear associations on the log-odds scale, and the chi-square test measures the dependence between each discretized feature and the target; both are straightforward to interpret. Lastly, XGBoost feature importance uses gradient boosting to evaluate feature importance based on each feature's contribution to the model's predictive performance [24].

For random forest and logistic regression, accuracy and feature-importance values were recorded in each fold of a stratified fivefold cross-validation (in each iteration, four folds were used for training and the remaining fold for validation), run 100 times (for logistic regression with a new random partition in each repetition, and for random forest with the same partition and a different random seed for the forest); for mutual information, the chi-square test and XGBoost, importance scores were computed on 1000 bootstrap resamples of the data. The feature-importance measures were method-specific: mutual information scores (k-nearest-neighbour estimator, $k = 3$), impurity-based importance for random forest, absolute coefficients of the standardized variables for logistic regression, chi-square statistics, and XGBoost importance scores. The mean and standard deviation of the importance values were computed across folds or resamples; the code that computes the importances, accuracy scores and visualizations of the five methods is provided in Supplementary Material S4, and the method-specific rankings of the re-estimation (Section 3.2) are reported in S5. To make the methods comparable, the vector of mean importances of each method was normalized to sum to one across features (relative importance), and the normalized scores were then averaged with equal weights across the five methods. The resulting mean normalized importance was used to rank the features (Fig. 1). Because this screening includes variables measured after graduation or concurrently with the outcome, it is an exploratory description of retrospective associations rather than a selection of predictors available at graduation.

3.2.2. Bayesian-optimized decision tree

Among the variables ranked in the previous analysis, we retained as predictors the nine variables related to education, skills and gender (gender, STEM degree, degree satisfaction, degree usefulness, ICT skills, and the theoretical, practical, language and ICT factors), five of which (degree usefulness and the four factors) refer to the period after graduation (Table 3), and excluded labour-market outcome variables such as first STEM employment, current employment, current salary, work schedule and professional status. We defined the modelling task as a binary classification problem in which the output was the predicted class of the target variable STEM occupation (employed in a STEM occupation or not). We then split the dataset into training and testing subsets with a 70/30 stratified split. This approach preserved the class proportions of the target variable in both subsets.

Decision trees are based on recursive partitioning and provide an interpretable modelling framework for classification problems [25]. We then used Bayesian optimization, implemented in the scikit-optimize library, to systematically search the hyperparameter space of a decision tree model. Bayesian optimization is particularly useful for hyperparameter tuning because it efficiently explores the search space by using previous evaluations to guide subsequent parameter choices [26].

The optimization process tuned parameters such as the splitting criterion, maximum tree depth, minimum number of samples per split or leaf, and cost-complexity pruning (alpha). Specifically, the Bayesian optimization was implemented using BayesSearchCV. The hyperparameter search space was defined as follows: splitting criterion {gini, entropy, log_loss}, maximum tree depth from 1 to 20, minimum number of samples required to split an internal node from 2 to 20, minimum number of samples required to be at a leaf node from 1 to 20, and cost-complexity pruning parameter ccp_alpha from 0.000 to 0.050, using a uniform prior. The optimization was performed over 20 iterations, using fivefold cross-validation and accuracy as the scoring metric.

After determining the optimal hyperparameter configuration, we used the scikit-learn library to retrain the decision tree model on the full training dataset and evaluated its performance on the test set. The evaluation included accuracy, precision, recall, F1-score and a confusion matrix. Precision, recall and F1-score are macro-averaged, i.e. the unweighted mean of the values for the STEM and non-STEM classes; we also report the values for the STEM (positive) class, the balanced accuracy, the area under the ROC curve (ROC-AUC) and the accuracy of a majority-class baseline that assigns all graduates to the class ‘not in a STEM occupation’ (0.804). Because the depth of this model made it difficult to interpret, we trained a second decision tree, restricting its depth to three while retaining the optimized values of the remaining hyperparameters. Fig. 2 shows this tree after cost-complexity pruning, which leaves a depth of two and four leaves; the splits removed by pruning do not change any predicted class, so the pruned and unpruned trees have identical class predictions, accuracy, precision, recall and F1-score and differ only in their predicted probabilities (ROC-AUC 0.837 vs. 0.857; Table 6). This pruned model aimed to strike a balance between interpretability and predictive performance. Evaluated with the same metrics, the pruned model has a slightly lower accuracy than the initial model but a much simpler structure (Section 4.1.2).

3.2.3. Association rule mining

We conducted an association rule analysis to identify how educational, demographic, and skill-related factors are associated with employment in a STEM occupation and, complementarily, with first STEM employment and current employment. Association rule mining is a data mining technique used to identify frequent co-occurring itemsets and implication rules in large datasets [27]. To prepare the data for this analysis, each category level of the variables included in the analysis was transformed into a separate binary item using one-hot encoding (e.g., ‘Gender = male’, ‘ICT factors = a lot’); missing and ‘don't know’ answers did not generate items. Frequent itemsets were then mined with the Apriori algorithm [27,28] implemented in Python (code in Supplementary Material S4).

Minimum support and minimum confidence define the thresholds for frequent itemsets and for the reliability of the generated rules, respectively. An exploratory grid search over lower thresholds (minimum support 0.001–0.01; minimum confidence 0.3–0.5) was used only to examine how the number of rules depends on these parameters, since lower thresholds necessarily generate more rules. The rules reported in Tables 4 and 5 use a minimum support of 0.02, so that each rule covers at least 634 graduates, and a minimum confidence of 0.5.

Association rules were then computed with the restriction that antecedents consist of items from the educational, demographic and skill-related variables (gender, STEM degree, degree satisfaction, degree usefulness, ICT skills and the theoretical, practical, language and ICT factors) and the consequent is the single item ‘STEM occupation = yes’. The resulting rules were ranked by lift, which quantifies how much more frequently antecedents and consequent co-occur than expected under independence (see Table 4). The analysis was repeated with antecedents restricted to the variables of sets A and B (Table 3). As a complementary analysis (Supplementary Material S5), we also mined rules whose consequents combine ‘STEM occupation = yes’, ‘first STEM employment = yes’ and ‘current employment = employed’; because every graduate employed in a STEM occupation is employed, rules in which

Table 4
Association rules with the highest lift whose consequent is STEM occupation = yes. Items are shown as variable = level (minimum support 0.02; minimum confidence 0.5).

Antecedents (consequent: STEM Occupation = yes)	Support	Confidence	Lift
Gender = male; Degree Usefulness = yes; STEM Degree = yes; Degree Satisfaction = yes; ICT Factors = a lot; ICT Skills = advanced	0.022	0.791	4.027
Gender = male; Degree Usefulness = yes; STEM Degree = yes; ICT Factors = a lot; ICT Skills = advanced	0.029	0.790	4.019
Gender = male; STEM Degree = yes; Degree Satisfaction = yes; ICT Factors = a lot; ICT Skills = advanced	0.024	0.788	4.011
Gender = male; Degree Usefulness = yes; STEM Degree = yes; ICT Factors = a lot; Practical Factors = a lot	0.020	0.785	3.998
Gender = male; STEM Degree = yes; ICT Factors = a lot; ICT Skills = advanced	0.033	0.782	3.979

Table 5
Impact of excluding each antecedent variable on the association rules whose consequent is STEM occupation = yes: number of rules, mean confidence, mean lift and maximum lift.

Excluded variable	Number of rules	Mean confidence	Mean lift	Maximum lift
None (all antecedents)	154	0.645	3.284	4.027
Gender	91	0.640	3.257	3.959
Degree Usefulness	81	0.636	3.238	4.011
STEM Degree	35	0.584	2.974	3.532
Degree Satisfaction	97	0.645	3.281	4.019
ICT Factors	103	0.625	3.181	3.895
ICT Skills	82	0.650	3.310	3.998
Theoretical Factors	136	0.644	3.281	4.027
Practical Factors	104	0.639	3.255	4.027
Language Factors	138	0.646	3.288	4.027

Table 6
Test-set performance of the decision trees by predictor set. Panel A: all graduates (test n = 9496); panel B: employed graduates only (test n = 8138). Precision, recall and F1 are macro-averaged unless stated otherwise. All re-estimations fit the preprocessing on the training data only; the training-only preprocessing re-estimation of set C retains the nine predictors originally selected after inspecting the full-data screening, whereas sets A and B are defined by temporal criteria. The training-only preprocessing re-estimation uses the original integer coding; the rows for sets A–D use the explicit coding (Section 3.2) unless labelled 'original coding'.

Predictor set	Model	Accuracy	Balanced accuracy	F1 (macro)	F1 (STEM class)	ROC-AUC
Panel A: all graduates (test n = 9496)						
Majority-class baseline	–	0.804	0.500	0.446	0.000	0.500
Main analysis, set C (Fig. 2)	Pruned tree (depth 2)	0.846	0.774	0.764	0.626	0.837
Training-only preprocessing re-estimation (original set C)	Bayesian-optimized	0.858	0.759	0.768	0.623	0.872
Training-only preprocessing re-estimation (original set C)	Depth-3	0.846	0.774	0.764	0.626	0.857
Set A, antecedent-only (before/at graduation)	Bayesian-optimized (one split)	0.825	0.805	0.760	0.634	0.805
Set B (A + skills/evaluations)	Bayesian-optimized	0.834	0.745	0.741	0.586	0.852
Set B (A + skills/evaluations)	Depth-3	0.833	0.662	0.686	0.472	0.842
Set C (main-tree predictors)	Bayesian-optimized	0.856	0.753	0.763	0.614	0.874
Set C (main-tree predictors)	Depth-3	0.855	0.768	0.769	0.629	0.859
Set D (all predictors)	Bayesian-optimized (depth 3)	0.912	0.840	0.854	0.763	0.910
Panel B: employed graduates only (test n = 8138)						
Majority-class baseline	–	0.771	0.500	0.435	0.000	0.500
Set A, antecedent-only (before/at graduation)	Bayesian-optimized (one split)	0.830	0.809	0.780	0.675	0.809
Set B (A + skills/evaluations)	Bayesian-optimized	0.832	0.806	0.781	0.674	0.847
Set C (main-tree predictors), original coding	Bayesian-optimized	0.840	0.774	0.774	0.652	0.858
Set D (all predictors)	Bayesian-optimized (one split)	0.900	0.834	0.851	0.765	0.834

'current employment = employed' accompanies 'STEM occupation = yes' are redundant and were removed.

Finally, we conducted a variable-exclusion analysis: by excluding each antecedent variable in turn, we measured how its absence affects the number of rules generated, their mean confidence, their mean lift and the maximum lift (see Table 5).

4. Results and discussion

4.1. Results

This subsection is divided into four parts: one for each of the three methods introduced in the methodology and a fourth one with the sensitivity analyses.

4.1.1. Feature selection

We distinguish two main types of variables associated with STEM occupations: labour-market (economic) variables and educational/demographic variables (see Fig. 1). First STEM employment ranks first, and other labour-market variables, such as current employment, work schedule, professional status and current salary, also rank among the ten most important variables. These variables, however, describe the employment trajectory itself and are measured after graduation or concurrently with the outcome (Table 3); their high importance therefore reflects their proximity to the outcome rather than predictive information available at graduation, and

they were not used as predictors in the decision tree. Among educational and demographic variables, STEM degree ranks highest (second overall), followed by ICT skills, gender and degree usefulness; the ICT factors, which refer to the current job, also rank high.

Labour-market variables describe the employment conditions of STEM jobs: among employed graduates, those in STEM occupations more often hold permanent contracts (67.5% vs. 54.4%) and full-time jobs (96.3% vs. 83.1%) and more often earn €2000 or more per month (31.4% vs. 19.2% of those reporting their salary). Educational and demographic variables capture the qualifications, skills and socio-demographic attributes associated with STEM careers. For instance, graduates with STEM degrees or advanced ICT skills are more likely to hold STEM-related positions, which is consistent with the technical content of these jobs. Degree usefulness reflects graduates' retrospective assessment of whether their degree helped them to find a job and may therefore partly reflect their labour-market experience. Gender is also associated with STEM employment, in line with the gender gaps in STEM study choices and STEM employment documented in Section 2.

4.1.2. *Bayesian-optimized decision tree*

The pruned decision tree (Fig. 2) uses three variables: STEM degree at the root, ICT skills in the branch of graduates without a STEM degree, and degree usefulness in the branch of graduates with a STEM degree. Graduates without a STEM degree are unlikely to hold STEM jobs (6.2% in the training set), although the proportion is higher among those with advanced ICT skills (15.3%) than among those with basic or intermediate skills (5.0%); this split refines the predicted probabilities but does not change the predicted class. Among graduates with a STEM degree, 60.4% of those who report that their degree helped them to find a job work in a STEM occupation, compared with 33.9% of those who report that it did not or who gave no valid answer. In this tree, only STEM degree and degree usefulness change the predicted class.

On the test set, the pruned tree achieves an accuracy of 0.846, compared with 0.804 for a majority-class baseline that assigns all graduates to the class 'not in a STEM occupation', and a balanced accuracy of 0.774. The best model found by Bayesian optimization achieves a slightly higher accuracy of 0.858. The macro-averaged metrics are comparable between the two models: the pruned model has a macro precision of 0.756 versus 0.777 for the Bayesian model, a macro recall of 0.774 versus 0.759, and a macro F1-score of 0.764 versus 0.768. For the STEM class, the pruned model has a precision of 0.598, a recall of 0.656 and an F1-score of 0.626 (ROC-AUC = 0.837). The confusion matrix of the explainable model is shown in Fig. 3.

4.1.3. *Association rule mining*

Table 4 lists the rules with the highest lift for employment in a STEM occupation. The rule with the highest lift (4.027) indicates that 79.1% (confidence = 0.791) of the male graduates with a STEM degree and advanced ICT skills who would study the same degree again, who report that their degree helped them to find a job and that ICT training contributed 'a lot' to obtaining their current job are employed in a STEM occupation, about four times the proportion in the whole sample (19.6%).

The other high-lift rules show slight variations (e.g., practical skills contributing 'a lot' to obtaining the current job instead of advanced ICT skills), but all of them include a STEM degree, male gender and a strong contribution of ICT training to obtaining the

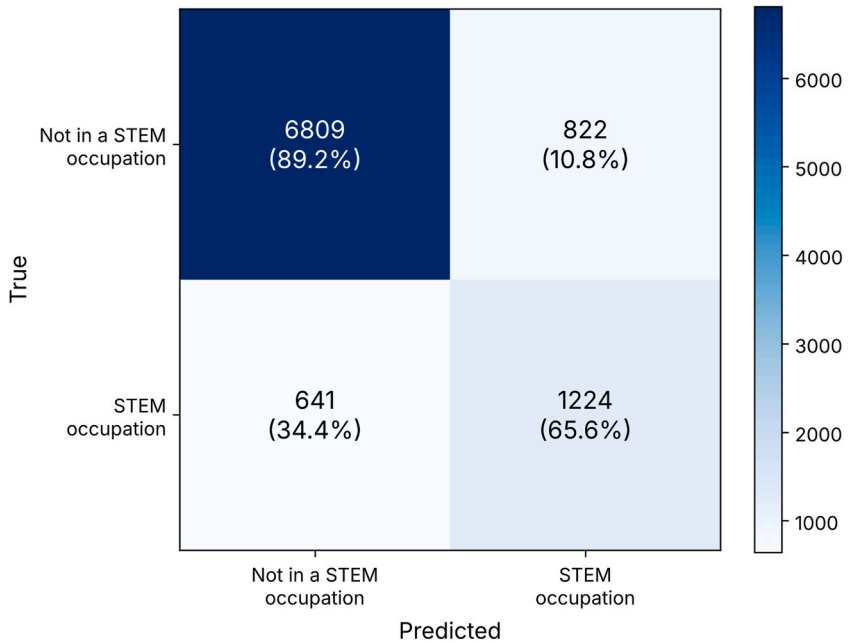


Fig. 3. Confusion matrix of the pruned decision tree model on the test set. The matrix compares the true and predicted classes of the binary target variable *STEM occupation*, where class 0 (not in a STEM occupation) corresponds to graduates employed in non-STEM occupations or not employed, and class 1 corresponds to graduates employed in STEM occupations. Cell values are reported as counts, with row percentages in parentheses.

current job. Because the ICT, practical, language and theoretical factors refer to the current job, rules containing them partly describe the characteristics of STEM jobs rather than antecedents of STEM employment. With antecedents restricted to the characteristics of set A, the strongest rule (lift = 3.045; confidence = 0.598) combines male gender, a STEM degree, a scholarship and having chosen the degree for its future employment prospects, and no combination of set A characteristics without a STEM degree reaches the minimum confidence of 0.5. When the skills and evaluations of set B are added, the strongest rule (lift = 3.603; confidence = 0.708) combines a STEM degree with advanced ICT skills, English as the main foreign language, no language certification and willingness to study the same degree again (Supplementary Material S5).

The variable-exclusion analysis (Table 5) confirms the central role of the STEM degree: excluding it reduces the number of rules from 154 to 35, the mean confidence from 0.645 to 0.584, the mean lift from 3.284 to 2.974 and the maximum lift from 4.027 to 3.532. Excluding the ICT factors produces the next largest reduction (mean lift 3.181; maximum lift 3.895), whereas excluding any other variable, including degree usefulness (mean lift 3.238; maximum lift 4.011), changes the mean lift by less than 0.05 and the maximum lift by less than 0.07. Because excluding a variable also removes its items, the number of rules decreases mechanically whenever a variable is excluded; the confidence- and lift-based measures are therefore more informative than the rule counts.

4.1.4. *Sensitivity analyses: training-only preprocessing re-estimation, antecedent-only predictors and employed graduates*

We first re-estimated the feature ranking and the decision trees with training-only preprocessing (Section 3.2), that is, with all preprocessing fitted on the training data only. With the original variables and coding, the re-estimated depth-three tree produces exactly the same predictions as the tree in Fig. 2 (accuracy 0.846; macro precision 0.756, recall 0.774 and F1-score 0.764), and the re-estimated Bayesian-optimized tree reaches an accuracy of 0.858 (macro precision 0.778, recall 0.759 and F1-score 0.768), practically identical to the values reported in Section 4.1.2 (Table 6). Using the explicit coding (Section 3.2) also leaves the results essentially unchanged (accuracy 0.856 for the Bayesian-optimized tree and 0.855 for the depth-three tree). Because these re-estimations keep the nine predictors that were originally selected after inspecting the full-data screening, they control for leakage in preprocessing and tuning but not in that selection step; the predictor sets defined by temporal criteria below provide checks that do not depend on it.

Second, we repeated the predictive analysis, with the explicit coding, using predictor sets defined by their timing (Table 3): set A includes only characteristics determined before or at graduation; set B adds the skills and degree evaluations reported in 2019 that do not refer to a specific job; set C is the set of nine predictors used in the main tree, which includes degree usefulness and the current-job factors; and set D includes all 32 predictors. The antecedent-only sensitivity analysis, which uses characteristics fixed before or at graduation (set A), confirms that holding a STEM degree is the dominant predictor of subsequent STEM employment (mean normalized importance 0.836), followed by gender (0.077). With set A, the Bayesian-optimized tree reduces to a single split on STEM degree, which correctly classifies 77.2% of the graduates employed in STEM occupations and 83.8% of the remaining graduates (accuracy 0.825, compared with 0.804 for the majority-class baseline; balanced accuracy 0.805). When age group, disability and country, which were recorded at the interview, are also removed, the tree and its predictions are identical, and STEM degree and gender remain the two most important variables (0.871 and 0.081). When the skills and evaluations reported in 2019 are included (set B), ICT skills become the second most important variable (0.116), followed by gender (0.052), and the Bayesian-optimized tree reaches an accuracy of 0.834 (balanced accuracy 0.745; ROC-AUC 0.852). However, because these skills were measured retrospectively, their values cannot be assumed to have been available at graduation.

Accuracy and ROC-AUC increase as later information is added (Table 6): relative to set A (accuracy 0.825), accuracy rises to 0.834 with the skills and evaluations reported in 2019 (set B), to 0.856 with the predictors of the main tree (set C; ROC-AUC 0.874) and to 0.912 with all predictors (set D; ROC-AUC 0.910). The gain of set D is mainly attributable to first STEM employment, which ranks first in this set (0.355), followed by STEM degree (0.136). Because the single split of set A assigns all graduates with a STEM degree to the STEM class, set A has a higher balanced accuracy (0.805) and STEM-class F1-score (0.634) than sets B and C, at the cost of lower accuracy and lower precision for the STEM class (0.538).

Finally, because the reference class includes graduates who were not employed at the interview, we repeated the sensitivity analyses with the 27,124 employed graduates only, so that the comparison is strictly between STEM and non-STEM employment (Table 6, panel B; Supplementary Material S5). The results are very similar: with set A, the tree again reduces to a single split on STEM degree (accuracy 0.830, compared with 0.771 for the majority-class baseline; balanced accuracy 0.809), and STEM degree remains by far the most important predictor (0.818), followed by gender (0.085); with set B, ICT skills rank second (0.106); the depth-three tree based on the predictors of the main analysis again splits first on STEM degree and then on ICT skills and degree usefulness; and the accuracy of set C (with the original coding) and set D is 0.840 and 0.900, respectively.

4.2. *Discussion*

The results obtained from the three methods employed (decision tree, feature selection, and association rules) consistently indicate that holding a STEM degree is the factor most strongly associated with employment in STEM occupations. In the decision tree, in addition, STEM graduates who consider that their degree helped them to find a job are more often employed in STEM occupations (60.4% vs. 33.9%). Because degree usefulness is a retrospective assessment reported about five years after graduation, this second association may partly reflect graduates' labour-market experience. The predominance of the STEM degree, by contrast, does not depend on post-graduation information: using only characteristics fixed at graduation, a single split on the STEM degree correctly classifies 77.2% of the graduates employed in STEM occupations and 83.8% of the others, and the same pattern holds among employed graduates only. Likewise, the gender variable, although absent from the pruned decision tree, is identified in feature selection, in the sensitivity analysis (where it is the second predictor among characteristics fixed at graduation) and in the association rules. In our data,

40.4% of men and 19.0% of women hold a STEM degree, and among STEM-degree holders 57.6% of men and 47.9% of women work in STEM occupations, a pattern consistent with the 'leaky pipeline' at labour-market entry documented for Spain [16]. Gender is therefore clearly associated with STEM employment, in line with previous studies (Section 2). This result should be interpreted mainly in relation to the persistent gender gap in STEM degree choice and subsequent labour-market entry, rather than as a direct causal effect of gender itself.

While previous studies have documented differences in labour-market outcomes by field of study and the persistence of gender inequalities in STEM pathways, our analysis shows that these patterns are also identified by interpretable machine learning methods applied to a large national survey of Spanish university graduates.

A first limitation is that the EILU does not record graduates' earlier relationship with mathematics, which has been linked to STEM career expectations. Using PISA 2022 data from four countries, Caspi and Gorsky [29] find that 9th–10th grade students' expectations of working in science, mathematics or engineering professions at age 30 are associated with their motivation to learn mathematics (interest, perceived importance of the subject and self-efficacy), which also mediates the associations of gender, disciplinary climate, teacher support and instructional strategies fostering conceptual understanding with these expectations. Similarly, in a sample of Chinese high-school students, Jiang et al. [30] find that self-efficacy and outcome expectations in STEM courses mediate the association between career awareness and STEM career interests.

A further limitation is that the available dataset does not include several contextual and psychological variables that may influence STEM career trajectories, such as family background, parental education, socioeconomic status, school processes, community environment, early exposure to STEM activities, role models, motivation, self-efficacy, or expectations about future career outcomes. The omission of these variables may limit the explanatory scope of the models and reinforce the need to interpret the findings as predictive and associative rather than causal. In addition, all information was collected retrospectively in a single survey wave, so that some skills and evaluations (e.g., ICT skills or degree satisfaction) were reported in 2019 and may have been influenced by post-graduation experiences.

Future research could triangulate these findings by combining national survey data with longitudinal educational records, school-level and family-background information, and qualitative evidence such as in-depth interviews. Such an approach would make it possible to better understand how STEM aspirations, educational choices, university experiences, and labour-market outcomes evolve over time, particularly among women and other groups underrepresented in STEM fields.

5. Conclusions

Decision tree–based models and association rules provide a complementary view of the factors associated with STEM labour market outcomes. The hierarchical analysis of decision trees highlights that holding a STEM degree is the variable most strongly associated with employment in this field and that, among STEM graduates, reporting that the degree helped to find a job is associated with a higher proportion in STEM occupations. In parallel, association rules underscore the importance of STEM education, gender, and ICT skills, among other factors. The predominance of the STEM degree, followed by gender, is robust to restricting the predictors to characteristics fixed at graduation and the sample to employed graduates. Most graduates employed in STEM occupations (77.2%) hold a STEM degree.

Moreover, the gender gap in STEM study choices [13,14] and in access to STEM positions [14,16] is widely documented, and it is also observed in the present data. Specific programmes have been implemented in Spain to increase women's participation in engineering and ICT. For example, since 2011 the Higher Technical School of Engineering of the University of Valencia (ETSE-UV) has run a programme to promote women's participation at bachelor's, master's and doctoral levels; female enrolment in its ICT degrees has remained above the Spanish average, which declined over the period, and an exploratory evaluation suggests that the programme has partly contributed to higher female graduation rates [31].

Although ICT skills and the language- and ICT-related factors also appear in the association rules, their contribution is smaller than that of the STEM degree (Tables 3 and 5). Country, type of university (public or private, in-person or distance), scholarship, studies abroad and language certification have low importance in the predictor-set analyses (Table 3), although scholarship and language certification appear in some high-lift rules together with a STEM degree (Section 4.1.3).

Moving forward, public policies and educational institutions could focus on early guidance toward STEM fields, particularly for women, and on the development of ICT skills, which, as reported in 2019, are associated with STEM employment even among graduates without a STEM degree (15.3% vs. 5.0%; Section 4.1.2), although the direction of this association cannot be established with our data. Degree satisfaction, measured as the willingness to study the same degree again, is relatively high: 72.0% of the graduates who would study at university again, the only ones asked this question, would choose the same degree (70.0% of STEM graduates and 72.8% of other graduates), which corresponds to 61.5% of all graduates; this variable showed little predictive value for STEM employment (Fig. 1 and Table 3). Strengthening the links between university programmes and employment opportunities may nevertheless help graduates, particularly those whose degree did not help them to find a job, to make better use of their training.

CRediT authorship contribution statement

Luis Ángel Calvo Pascual: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Ana Zapatero González:** Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Conceptualization.

Data availability statement

Data included in article/supplementary material is referenced in the article. The analysis uses publicly available, anonymised microdata of the 2019 Survey on the Labour Market Insertion of University Graduates (EILU 2019) released by the Spanish National Statistics Institute (INE) [12]. The processed analytical dataset and its codebook (Supplementary Material S6) and all the code used in the analysis (Supplementary Material S4) are provided as supplementary material.

Ethics statement

Review and/or approval by an ethics committee was not needed for this study because it is a secondary analysis of publicly available, fully anonymised official statistical microdata released by the Spanish National Statistics Institute (INE). Informed consent was not required for this study because the data were collected by INE under the Spanish statistical legislation and released as anonymised microdata that do not allow the identification of individual respondents.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used Claude (Anthropic) to improve the writing and the fluency of the text. The authors reviewed and edited the output as needed and take full responsibility for the content of the published article.

Funding declaration

This work was supported by Amazon through a research agreement.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank the Telling and Code teams for the fruitful discussions on the topic.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2026.e45629.

References

[1] B.M. Torres-Olave, STEM Employment in the New Economy: a Labor Market Segmentation Approach, The University of Arizona, 2013. https://hdl.handle.net/10150/305866.
[2] C.L. Garrouste, M. Rodrigues, Employability of young graduates in Europe, Int. J. Manpow. 35 (4) (2014) 425–447, https://doi.org/10.1108/IJM-05-2013-0106.
[3] M. Branine, A. Avramenko, A comparative analysis of graduate employment prospects in European labour markets: a study of graduate recruitment in four countries, High. Educ. Q. 69 (4) (2015) 342–365, https://doi.org/10.1111/hequ.12076.
[4] F. Green, G. Henseke, Europe's evolving graduate labour markets: supply, demand, underemployment and pay, Journal for Labour Market Research 55 (2021) 2, https://doi.org/10.1186/s12651-021-00288-y.
[5] M.J. Freire-Seoane, B. López-Bermúdez, C. Pais-Montes, Employability profiles: the case of a Spanish university, Cuad. Econ. 43 (123) (2020) 341–352, https://doi.org/10.32826/cude.v43i123.238.
[6] M. Bacovic, Z. Andrijasevic, B. Pejovic, STEM education and growth in Europe, J. Knowl. Econ. 13 (3) (2022) 2348–2371, https://doi.org/10.1007/s13132-021-00817-7.
[7] M.J. López López, P. Simó, J. Marco, Understanding STEM career choices: a systematic mapping, Heliyon 9 (6) (2023) e16676, https://doi.org/10.1016/j.heliyon.2023.e16676.
[8] E.C. Meroni, E. Vera-Toscano, The persistence of overeducation among recent graduates, Labour Econ. 48 (2017) 120–143, https://doi.org/10.1016/j.labeco.2017.07.002.
[9] S. Budría, A. Moro-Egido, Overqualification, skill mismatches and wages in private sector employment in Europe, Technol. Econ. Dev. Econ. 20 (3) (2014) 457–483, https://doi.org/10.3846/20294913.2014.883341.
[10] D. Verhaest, S. Sellami, R. Van Der Velden, Differences in horizontal and vertical mismatches across countries and fields of study, Int. Labour Rev. 156 (1) (2017) 1–23, https://doi.org/10.1111/j.1564-913X.2015.00031.x.
[11] A. Lacuesta, M. Martínez-Matute, J. Sainz, I. Sanz, Desajustes entre demanda y oferta de titulaciones en las universidades públicas presenciales, Papeles Econ. Espanola 180 (2024) 98–113.
[12] National Statistics Institute (INE), Encuesta De Inserción Laboral De Titulados Universitarios (EILU) 2019 [Survey on the Labour Market Insertion of University Graduates 2019], INE, Madrid, 2020. https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176991&menu=resultados&secc=1254736195339&idp=1254735573113 (press release: https://www.ine.es/prensa/eilu_2019.pdf.
[13] A.M. González Cervera, Y. González Arechavala, Informe Estudios STEM en la Unión Europea y participación de la mujer. http://hdl.handle.net/11531/84866, 2023.

[14] C. Tarín Quirós, E. Guerra Morales, R. Rivera Pastor, A. Fraile Carmona, M. Sáinz, U. Madinaveitia Herrera, Women in the Digital Age, European Commission, 2018, https://doi.org/10.2759/526938.
[15] Y.J. Xu, Attrition of women in STEM: examining job/major congruence in the career choices of college graduates, J. Career Dev. 44 (1) (2017) 3–19, https://doi.org/10.1177/0894845316633787.
[16] M.T. Valdés, H. Solga, The STEM leaky pipeline at labor market entry in Spain: the role of job competition and social origin, Soc. Sci. Res. 124 (2024) 103092, https://doi.org/10.1016/j.ssresearch.2024.103092.
[17] S. González-Pérez, R. Mateos de Cabo, M. Sáinz, Girls in STEM: is it a female role-model thing? Front. Psychol. 11 (2020) 2204, https://doi.org/10.3389/fpsyg.2020.02204.
[18] J.L. Cleary, M.R. Kerrigan, M. Van Noy, Towards a new understanding of labor market alignment, in: M.B. Paulsen (Ed.), Higher Education: Handbook of Theory and Research, vol. 32, Springer, 2017, pp. 577–629, https://doi.org/10.1007/978-3-319-48983-4_12.
[19] UNESCO Institute for Statistics, ISCED Fields of Education and Training 2013 (ISCED-F 2013), UNESCO Institute for Statistics, Montreal, 2014. ISBN 978-92-9189-150-4.
[20] R. Noonan, *STEM Jobs: 2017 Update* (ESA Issue Brief #02-17). Office of the Chief Economist, Economics and Statistics Administration, U.S. Department of Commerce, 2017.
[21] G. Siekmann, P. Korbel, *Defining 'STEM' Skills: Review and Synthesis of the Literature* (Support Document 1), National Centre for Vocational Education Research (NCVER), Adelaide, 2016.
[22] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182, https://doi.org/10.5555/944919.944968.
[23] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[24] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.
[25] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Routledge, 2017, https://doi.org/10.1201/9781315139470.
[26] J. Snoek, H. Larochelle, R.P. Adams, Practical bayesian optimization of machine learning algorithms, Adv. Neural Inf. Process. Syst. 25 (2012) 2951–2959.
[27] R. Agrawal, T. Imieliński, A. Swami, Mining association rules between sets of items in large databases, in: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 1993, pp. 207–216, https://doi.org/10.1145/170035.170072.
[28] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proceedings of the 20th International Conference on Very Large Data Bases, 1994, pp. 487–499.
[29] A. Caspi, P. Gorsky, STEM career expectations across four diverse countries: motivation to learn mathematics mediates the effects of gender and math classroom environments, International Journal of STEM Education 11 (2024) 52, https://doi.org/10.1186/s40594-024-00511-5.
[30] H. Jiang, L. Zhang, W. Zhang, Influence of career awareness on STEM career interests: examining the roles of self-efficacy, outcome expectations, and gender, International Journal of STEM Education 11 (2024) 22, https://doi.org/10.1186/s40594-024-00482-7.
[31] C. Botella, S. Rueda, E. López-Iñesta, P. Marzal, Gender diversity in STEM disciplines: a multiple factor problem, Entropy 21 (1) (2019) 30, https://doi.org/10.3390/e21010030.