IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received March 19, 2022, accepted April 17, 2022, date of publication April 25, 2022, date of current version May 5, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3170421

Predicting Student Employability Through the Internship Context Using Gradient Boosting Models

OUMAIMA SAIDANI, (Member, IEEE), LEILA JAMEL MENZLI, AMEL KSIBI, NAZIK ALTURKI, AND ALA SALEH ALLUHAIDAN

Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia

Corresponding author: Nazik Alturki (namalturki@pnu.edu.sa)

The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number PNU-DRI-RI-20-003.

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT Universities around the world are keen to develop study plans that will guide their graduates to success in the job market. The internship course is one of the most significant courses that provides an experiential opportunity for students to apply knowledge and to prepare to start a professional career. However, internships do not guarantee employability, especially when a graduate's internship performance is not satisfactory and the internship requirements are not met. Many factors contribute to this issue making the prediction of employability an important challenge for researchers in the higher education field. In this paper, our aim is to introduce an effective method to predict student employability based on context and using Gradient Boosting classifiers. Our contributions consist of harnessing the power of gradient boosting algorithms to perform context-aware employability status prediction processes. Student employability prediction relies on identifying the most predictive features impacting the hiring opportunity of graduates. Hence, we define two context models, which are the student context based on the student features and the internship context based on internship features. Experiments are conducted using three gradient boosting classifiers: e**X**treme **G**radient **B**oosting (XGBoost), **C**ategory **B**oosting (CatBoost) and **L**ight **G**radient **B**oosted **M**achine (LGBM). The results obtained showed that applying LGBM classifiers over the internship context performs the best compared to student context. Therefore, this study provides strong evidence that the employability of graduates is predictable from the internship context.

INDEX TERMS Internship, employability, context awareness, machine learning, prediction, features.

I. INTRODUCTION

Currently, in most universities, internship programs are considered a mandatory part of the student curriculum. The internship is defined as "a short-term practical work experience in which students receive training and gain experience in a specific field or career area of their interest" [1]. Internship programs ease the transition from universities to the workplace, as they are an effective method to fill in the gap between what is learned in universities and employment demands [2]. An internship program provides a student with the opportunity to practice the content learned during lectures [3], integrate theoretical knowledge with practical experiences gained through experiential learning [4], become familiarized with the workplace [5], [6], clarify career expectations, and develop valuable practical experience and job-relevant skills [7]. Moreover, internship programs have proven to be one of the most important experiential learning activities that enhances graduates' employability [8]–[11]. However, planning, designing and coordinating an effective and successful internship is challenging [1], [12]. As a result, universities continue to produce graduates who may be considered unfit for the job market.

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

Although there are studies that focus on predicting employability, such as [13], due to the lack of internship data and the difficulty of gathering and analyzing these data, there is a shortage of research on the internship factors that affect employability of students after graduation [14]–[16].

Very limited research, such as [17] and [18], has been carried out to examine the impact of internship programs on the labor market opportunities [19]. There are even fewer studies that predict student employability based on their internship program experience [19]–[21].

Therefore, in this study, we aim to predict student employability based on their internship program and the context-related knowledge; to our knowledge, there is no prior research that has predicted student employability based on contextual knowledge.

Previous researchers have shed light on the importance of context awareness [22], [23]. However, 'context' meaning differs across studies, and there is no real consensus on the definition of the concept. Dey *et al.* [24] defined context as ''any information that can be used to characterize the situation of entities that are considered relevant to the interaction between a user and an application, including the user and the application themselves''. Other studies consider context as a set of information variables or attributes that have an important impact on decision-making [25]–[28]. In this study, we identify internship context as all the internship attributes that affect student employability. Similarly, the student context is defined as all the student attributes that affect student employability.

In the field of Machine Learning (ML), context awareness is a relatively new field of research. Researchers exploring ML and considering the context are very limited. In fact, [29] is the only study identified that focused on this aspect. The researchers compared the use of a trained general model that uses all contexts in contrast to a system made of a set of specialized models trained for each specific operating context. Moreover, to our knowledge, there is no research dealing with predicting student employability that consider contextual knowledge.

Process flexibility has been the focus of much research in many areas [26], [30]–[32]. It means fast reactivity to internal and external changes [30] and reflects the ease of evolving process models. In general, ML prediction processes are formalized as a set of activities and their relationships. We argue that introducing flexibility in the achievement of the different steps of the prediction process could be more suitable in the current changing atmosphere.

Therefore, in this work, we utilize Gradient Boosting Models to predict student employability through student and internship contexts and reveal the most predictive features leading to improved chances in employment.

Our contributions include mainly: (i) a flexible context-aware employability prediction process model, (ii) a context model for employability prediction, and (iii) a context-based M approach for student employability prediction. The output feature of this approach is the student employability status (i.e., identifying whether the student is most likely to be employed, unemployed, or continue his or her studies, or in training). Moreover, this approach determines the sensitive features, i.e., the most predictive features among the inputs.

The research questions for this study are as follows:

- RQ1: What are the best ML algorithms for employability prediction?

- RQ2: What are the most sensitive features that impact employability prediction?

- RQ3: Do student context and internship context influence employability prediction?

The rest of the paper is structured as follows: Section II presents a literature review that examines relevant studies and sets the conceptual framework for this work. In section III, the methodology steps are presented, where the context-aware employability process is explained. Section IV details the experiments and results. Section V outlines the discussion. Finally, section VI concludes this work with some limitations and improvement windows.

II. LITERATURE REVIEW

A. INTERNSHIP AND EMPLOYMENT

A well planned internship program, jointly coordinated by industry representatives and academic institutions, is expected to increase its effectiveness. An internship program is successful whenever students are satisfied, assigned tasks are relevant and supervisors are qualified [1], [6], [12], [33]–[35]. Furthermore, some studies identified the correlation between internship satisfaction and employability, i.e., consultation and treatment provided during internship influence the satisfaction of students with the internship and their decision to remain in the same industry in the future [33], [36], [37].

Eurico *et al.* [38] claim that the satisfaction of a student with the internship program enhances his or her employability skills. Furthermore, Hugo [19] assumes that internships, majors and extracurricular activities impact employability.

Previous studies have proven that internship programs generally increase employability [9]–[11]; however, it is still unclear which internship features play a major role in employability.

B. ML FOR EMPLOYMENT PREDICTION

Machine learning is widely applied in many research fields [39], [40]. In higher education, ML is used mainly to enhance curriculum outcomes and graduate features. Many researchers are interested in this field and have conducted several studies to discuss the contribution of ML in continuous quality improvement.

As recommended by the research community [41], [42], we used Google Scholar, IEEE Xplore, ACM Digital Library, ScienceDirect and Scopus Database to extract relevant published papers between 2012 and 2021.

The literature search was based on many fields: employability in general, graduate employability in higher education,

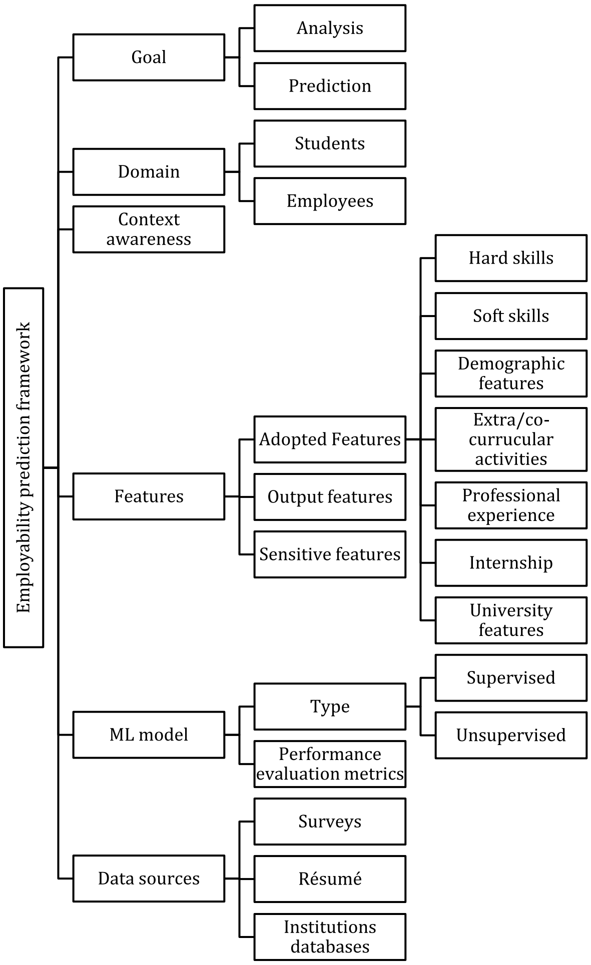

FIGURE 1. **Literature review framework.**

prediction, ML and ML algorithms. Twenty relevant studies were retained. Our examination of papers is conducted by the following criteria:

- *Goal*. The goal of the study.

- *Domain*. The application domain: who is the subject for which employability is predicted (i.e., employees, students)?

- *Context awareness*. Indicates whether the context is supported.

- *Adopted features*. The employability signals adopted in the study, e.g. computational skills, communication skills, number of majors, grades.

- *Output features*. The predicted outputs.

- *Sensitive features*. The selected features that truly influence the employability aspects.

- *ML model*. The ML model used.

- *Dataset sources*. The dataset source(s) used.

- *Performance evaluation metrics*. The performance evaluation metrics considered.

The following subsections analyze and compare the surveyed studies with respect to the framework proposed in Fig.1.

1) GOAL, DOMAIN AND CONTEXT AWARENESS

This subsection analyses the surveyed studies according to the following criteria: Goal, Domain and Context-awareness. Table 1 gives a summary of the retained studies.

a) Goal: Regarding the goal criterion, most of the studies aim to predict student employability. In the same vein, the study of L. Hugo [19] aimed to predict which students will graduate with a non-employment offer. Kommina *et al.* [43] and Kalpana and Venkatalakshmi [44] aimed to predict academic performance. Casuat *et al.* [45] focused on the identification of the most dominant attributes that affect employability. Patel *et al.* [46] aimed to predict the most suitable job domains.

b) Domain: with respect to the domain criterion, the analysis of the data gathered from the surveyed studies shows that nineteen of the proposed studies focus on student employability and performance prediction, and only one study worked on candidate employability prediction in general.

c) *Context awareness*: with respect to the context awareness criterion, none of the surveyed studies addressed contextual aspects or supported employability prediction based on context-related knowledge.

2) FEATURES

This subsection analyses the surveyed studies according to the following criteria: Adopted features, Output features, Sensitive features and Techniques used for determining the sensitive features (see Table 2 and Table 3).

Features are identified as factors influencing the success of internship programs. We distinguished three categories of features as follows.

a) *Adopted features*: we classified the different adopted features into six categories: *Hard skills*, *Soft skills*, *Demographic features*, *Extracurricular/co-curricular activities*, *Professional experience* and *Internship*. Table 2 represents the retained works according to these categories. Most of the studies focus mainly on Grade Point Average (GPA) [19], [44], [45], [47]–[51], assignments and exams [43], majors [19], student performances on technical, computational and analytical courses [44], [46], [48], [51]–[54].

Kommina *et al.* [43], Casuat *et al.* [45], Bai and Hira [48], Giri *et al.* [52] and Kumar and Babu [55] focus on soft skills. Personal and socio-demographic variables are adopted features in [19], [44], [45], [47], [48], [51], [53], [55], [56]. University related features are adopted in [20], [44], [48], [49], [55], [57].

Only a minority of studies support extracurricular and curricular activities [19], [20], [49], [50], [51], [55]. In the same vein, a few studies supported internship related features such as experience in internships [20], [21], or the number of internships [19]. Similarly, professional experience, such as “Number of companies worked previously” is supported only in [56].

TABLE 1. Related works: Goal and domain and context awareness.

Study and authors	Year	Goal	Domain	Context awareness
Hugo [19]	2018	Predict the students' career outcomes.	Students	Not supported
Othman et al. [20]	2018	Identify the factors that influence graduates employability.	Students	Not supported
Nunley et al. [21]	2016	Estimate the impact of college majors and internship experience on employment prospects.	Students	Not supported
Kommina et al. [43]	2020	Predict the students' academic performance and employability chances.	Students	Not supported
Kalpana and Venkatalakshmi [44]	2014	Analyze the students' performances.	Students	Not supported
Casuat et al. [45]	2020	Predict the students' employability.	Students	Not supported
Patel et al. [46]	2020	Predict the suitable students' job domains.	Students	Not supported
Casuat et al.[47]	2020	Identify the most predictive attributes among employability signals.	Students	Not supported
Bai and Hira. [48]	2021	Predict the students' employability.	Students	Not supported
Dubey and Mani [49]	2019	Predict the employability of high school students with local businesses for part-time jobs.	Students	Not supported
Pinto [50]	2019	Analyze the influence of academic performance and extracurricular activities on the perceived employability of students.	Students	Not supported
Jantawan and Tsai [51]	2013	Predict students' employability	Students	Not supported
Giri et al. [52]	2016	Predict the probability of an undergraduate student getting placed in an IT company.	Students	Not supported
Osmanbegovic and Suljic [53]	2012	Predict the students' success.	Students	Not supported
Laddha [54]	2021	Predict students' employability based on Technical Skills.	Students	Not supported
Kumar and Babu [55]	2019	Predict the students' employability.	Students	Not supported
Reddy et al. [56]	2021	Predict joining efficient candidates.	Employees	Not supported
Aviso et al. [57]	2020	Predict the employability of chemical engineering graduates based on UK university rankings.	Students	Not supported
Maheswari [58]	2020	Predict the student's performance in placement.	Students	Not supported
Almutairi [59]	2018	Predict the suitability of Information Systems' graduates.	Students	Not supported

b) Output features: eleven of twenty studies focused on placement [43], [58], [52], employability [19], [38], [45], [48], [51], hiring [49], recruitment [56], getting a job [55] and working [20]. Rare are those studies that concentrated on employability rate [21], company [58] or graduation [43]. Moreover, [59] focused on predicting the student matching to skills required by the Saudi industry.

c) Sensitive features are the selected features. Most of the studies did not identify the most sensitive features. Internship is considered as sensitive in only three studies [19]–[21]. In addition, internship is considered in [19] as the most sensitive variable, followed by specific majors and co-curricular activities. Extracurricular activities are considered in [50] as a sensitive feature. Moreover, mental alertness, manner of speaking, ability to express ideas and self-confidence are sensitive features in [47]. In [55], the most predictive features are as follows: aptitude and reasoning skills, communication skills, family income status, mentor and quality of teaching in the college.

d) Techniques for determining sensitive features: a number of techniques are used to determine the sensitive features such as Univariate Feature Selection technique, Recursive Feature Elimination technique, and Principal component Analysis technique [47], Logistic regression (LR) - P-values [19], Pearson correlation method and Kandel correlation method [54], and WEKA feature selection [20].

TABLE 2. Related works: Adopted features categories.

Study	Hard skills	Soft skills	Demographic features	Extra/co-curricular activities	Professional experience	University features	Internship
[19]	▪		▪	▪			▪
[20]	▪	▪	▪	▪		▪	▪
[21]	▪						▪
[43]	▪						
[44]	▪		▪			▪	
[45]	▪	▪	▪				
[46]	▪						
[47]	▪	▪	▪				
[48]	▪	▪	▪			▪	
[49]	▪		▪	▪			
[50]	▪		▪	▪	▪		
[51]	▪		▪		▪	▪	
[52]	▪	▪					
[53]	▪		▪				
[54]	▪						
[55]	▪	▪	▪	▪		▪	
[56]	▪		▪		▪		
[57]						▪	
[58]	▪	▪					
[59]	▪	▪					

TABLE 3. Related works: adopted features, output features and sensitive features.

Study	Adopted features	Output features	Sensitive features	
			Features	Technique used
[19]	GPA, major, number of co-curricular activities, number of internships, sex, ethnicity, international status, graduation date and number of majors.	Employability: {Employed, Not employed}.	Major, number of internships and co-curriculum activities.	Neural network (NN), LR (p-values).
[20]	Gender, age, race, state, faculty, field. CGPA, family income, co-curriculum, marital status, industrial internship, join any entrepreneurship programs, Bahasa Melayu skill, English language skill, interpersonal skill, critical and creative thinking, problem solving skill, analytical skill, teamwork, positive values, general knowledge, current status and financial assistance.	Employability: {Working, Not working}	Age, faculty	WEKA feature selection
			Age, faculty, field, family income, co-curriculum, marital status and English language skill.	Expert's opinion, feature selection and past research.
[21]	College major and internship experience.	Employability rate.		Regression model.
[43]	Student attendance, class test, laboratory externals, laboratory internals, assignments, final exams and midterm exams.	Graduation: {Pass, Not pass}. Placement: {Placed, Not placed}.	NA	NA
[44]	Classroom/group, number of students in group, attendance during morning/evening sessions, number of friends, number of hours spent studying daily, methods of study used, place used for studying, age, sex, previous school, type of school, type of secondary school, GPA, scores in: math 1, physics 1, social science 1, humanities 1, writing and reading 1, English 1 and computer 1.	NA	NA	NA
[45]	General appearance, manner of speaking, physical conditions, mental alertness, self-confidence, ability to present ideas, communication skills, students' performance rating, GPA and student program.	Employability: {Employable, Less employable}.	NA	NA
[46]	Computational skills.	Preferred work domains.	NA	NA
[47]	General appearance, manner of speaking, physical conditions, mental alertness, self-confidence, ability to present ideas, communication skills, students' performance rating, and GPA.	Employability: {Employable, Less employable}.	Mental alertness and manner of speaking.	Univariate feature selection technique.
			Mental alertness, Manner of speaking and ability to present ideas.	Recursive feature elimination technique.
			Self-confidence and manner of speaking.	Principal component analysis technique.
[48]	Age, stream, degree, department, university, CGPA, course type, gap between CG and PG, sex, category, city, state, known languages, communication skill, leadership, risk management, team management, problem solving skill, research ability, negotiation, programming skill, reasoning ability, quantitative skill and cognitive skill.	Employability: {Employed, unemployed}	NA	NA
[49]	GPA, school, grade level, social and task skills, available days to work and number of courses taken.	Hiring: {Hired, Not hired}.	NA	NA
[50]	GPA, Age, gender, Education, extracurricular activities and recruiting experience.	Employability: {Employed, Unemployed}.	Academic performance and extracurricular activities.	Multivariate analyses of variance.
[51]	Gender, province, degree, educational background, faculty, GPA, status and position.	Employability: Employed, Unemployed, Undetermined situation}	NA	NA
[52]	Technical skills, communication skills, analytical skills and teamwork.	Placement: {Placed, Not placed}.	NA	NA
[53]	GPA, passing grade on the exam, socio-demographic variables, achieved results from high school and from the entrance exam and attitudes towards studying.	Success: {Pass, Fail}.	GPA, entrance exam, study material, average weekly hours devoted to studying.	Chi-square test, One R-test, Info Gain test and Gain Ratio test.

TABLE 3. *(Continued.)* Related works: adopted features, output features and sensitive features.

[54]	Students' performance in various technical courses (C programming, data structures and algorithms, mobile app development, ML and web technology).	Placement: {Yes, No}.	C programming, data structure and algorithms, and MLgrades.	Pearson correlation method.
			C programming, data structure and algorithms grades.	Kandel correlation method.
[55]	Gender and caste, family income status, qualification and employment type of mother and father, high school board, medium of school, mentor, students SSC%, Students Inter%, preference for opting a certain course, students allotment of seat, communication skills, aptitude and reasoning skills, student's stay, number of study hours per day, attendance, quality of teaching in college, hours spent on social media, watching movies and porn and playing games, changing out with friends and use of a vehicle.	Getting job: {Yes, No}	NA	NA
[56]	Gender, age, marital status, years of experience, time in current company in months, number of companies worked previously, salary hike, education background, job description, compensation, years of promotion and distance from home.	Recruitment. {Join, Not join}.	NA	NA
[57]	University features: entry standards, research intensity, staff to student ratio, budget per student, and research quality.	Employability: {Yes, No}	NA	NA
[58]	10th percentage, 12th percentage, CGPA, personal skill and arrears.	Placement: {Placed, not placed} and company.	NA	NA
[59]	Skills, self-regulated learning and academic achievement of IS students.	The students' matching to skills required by industry.	NA	NA

3) ML MODEL, DATA SOURCES AND PERFORMANCE EVALUATION METRICS

This subsection analyzes the surveyed studies according to the criteria *ML model*, *Data sources* and *Performance* evaluation metrics, in accordance with the research questions posed as shown in Table 4.

a) *Machine Learning Model:* to predict employability of students or employees, some researchers used traditional approaches (i.e., statistical sampling, surveys) to predict employability rates [19]. In contrast, the majority employed ML algorithms as they show their high prediction performances. Indeed, we notice that approximately 95% of related works implemented or/and sometimes went even further to compare supervised ML models [19], [21], [44], [45], [47], [49], [51]–[56], [58], [59].

Limited studies have implemented unsupervised ML models [20], [43], [45], [57]; the authors of [48] proposed a hybrid model of a deep belief network and soft max regression (DBN-SR).

b) *Dataset Sources*: three types of dataset sources are used in the studies examined:

- Database: most studies exploited databases from registration units or/and institutional departments [20], [43]–[46], [51], [52], [54], [57].

- Surveys: surveys of students/employees and employers [44], [48], [55], [56], [58], [59].

- Databases and surveys: [19], [44], [47], [53].

We notice that few studies employed résumés ([21] and [50]) regardless of the difficulty of analyzing résumés and the scarcity of real résumés.

c) *Performance Evaluation Metrics*: most retained articles used one or many performance evaluation metrics to compare ML algorithms (i.e., precision, recall, FI score, accuracy, etc.). Only [21], [44], [46], [50] did not apply any metric.

4) DISCUSSION ON RESEARCH QUESTIONS

The previous comparative analysis provides evidence of a number of commonalities and limitations of the above studies.

First, with most of the studies, the output features are binary and allow predicting whether the student will have a job. The development of methods that offer multiple output values in predicting employability may prove most useful. However, we observe a lack of research supporting this possibility. With respect to this limitation, our work supports the prediction of employment status, i.e., whether the student will be employed, unemployed, continue studies, or go on training.

Second, none of the studies consider context knowledge in the prediction process. We also observe a lack of studies that focus on the contextual information that could have an impact on the prediction of the output features. The studies

TABLE 4. Related works: ML model, dataset sources and performance evaluation metrics.

Study	ML model	Dataset sources	Performance evaluation metrics
[19]	Supervised: Support Vector Machine (SVM), Nearest Neighbor (NN), LR, Discriminant analysis and Decision Tree (DT).	Institution databases and student surveys.	Accuracy, error, specificity, sensitivity and precision.
[20]	Supervised: DT, ANN and SVM.	Institution databases.	Accuracy, ROC, RMSE and the time taken to build the model.
[21]	Supervised: Regression model.	Résumés.	NA
[43]	Supervised: Hybrid Linear Vector Quantization (HLVQ).	Institution databases.	Precision, recall, F1 Score and accuracy.
[44]	Unsupervised: Clustering methods (density based, centroid based and distribution based).	Institution databases and student surveys.	NA
[45]	Supervised: DT, Random Forest (RF), SVM, K- Nearest Neighbor (KNN) and LR.	Institution databases.	Accuracy, precision, F1 score and recall metrics.
[46]	Unsupervised: NN (Keras).	Institution databases.	NA
[47]	Supervised: SVM, DT and RF.	Institution databases and graduate surveys.	Accuracy, precision, F1 score and recall metrics.
[48]	Supervised: Softmax regression. Unsupervised: Deep belief network.	Student surveys.	Accuracy, precision, sensitivity, kappa coefficient, specificity, F-score and prediction error value.
[49]	Supervised: LR, DT, RF, KNN and SVM.	Student surveys.	Precision, recall, F1 Score, accuracy, AUC (Area Under the Curve) and ROC (Receiver Operating Characteristics)
[50]	ML is not applied (an online survey is used: statistical sampling).	Résumés	NA
[51]	Supervised: Bayesian method, and Tree method.	Institution Database.	Accuracy, error and rate.
[52]	Supervised: KNN, SVM and LR.	Institution databases.	Confusion matrix and accuracy.
[53]	Supervised: Bayesian Classifier, NN and DT.	Institutions databases and student surveys.	Prediction accuracy, learning time and error rate.
[54]	Supervised: SVM, Naive Bayes (NB), LR, DT, RF, Ada-Boost and ANN.	Institution databases.	Accuracy, precision, recall and F1 Score.
[55]	Supervised: DT, Gaussian Naive Bayes (GaussianNB), SVM and KNN.	Student surveys.	Precision, recall, F1 Score and accuracy.
[56]	Supervised: DT, RF, GaussianNB and NN.	Employee surveys.	Accuracy, precision and recall.
[57]	Supervised: Rule-based Hyperbox model [18].	Institution databases.	Balanced accuracy, specificity and sensitivity.
[58]	Supervised: LR.	Student surveys.	Accuracy, precision and recall.
[59]	Supervised: DT, NB and NN.	Student surveys and recruiter surveys.	TP Rate, FP Rate, precision, recall and F-Measure and ROC Area (Receiver Operating Characteristics).

do not address the question of which contextual information can offer a prediction with better performance with respect to accuracy, precision, recall and F-score. In other words, the relationship between context and prediction has been neglected thus far by previous studies. This observation calls for the development of research that considers contextual information during the prediction process. We propose in this work to model and to include the context related to the student and to the internship in the prediction process.

III. METHODOLOGY : A CONTEXT-AWARE EMPLOYABILITY PREDICTION PROCESS

The contextual information characterizing the student profile and the internship is of great interest in the prediction process. This fact requires identifying solutions that allow context-based prediction. This study presents a new approach for predicting graduate employability status, based on contextual information related to the student profile and the internship. A large number of attributes and features are used by ML-based algorithms in this regard.

To enhance the functionality and the performance of the latter, a context-aware ML-based model is defined. In general, a context-aware system supports the idea of using available information for specific user needs to improve the behavior of the system itself. We extend this idea to this work: the proposed ML model considers the student profile and the internship conditions. For this purpose, a "context-aware feature selection" step is added to the standard ML process to decide which relevant and sensitive features (data) to use for learning and testing the model.

Even better, we think that representing the different alternatives for achieving the rest of the steps (i.e., data collecting, feature engineering, model training, model evaluating and model employing) is of great help to follow the process fulfillment (i.e., intentions) and allow flexibility by using alternatives (techniques). Furthermore, ML processes

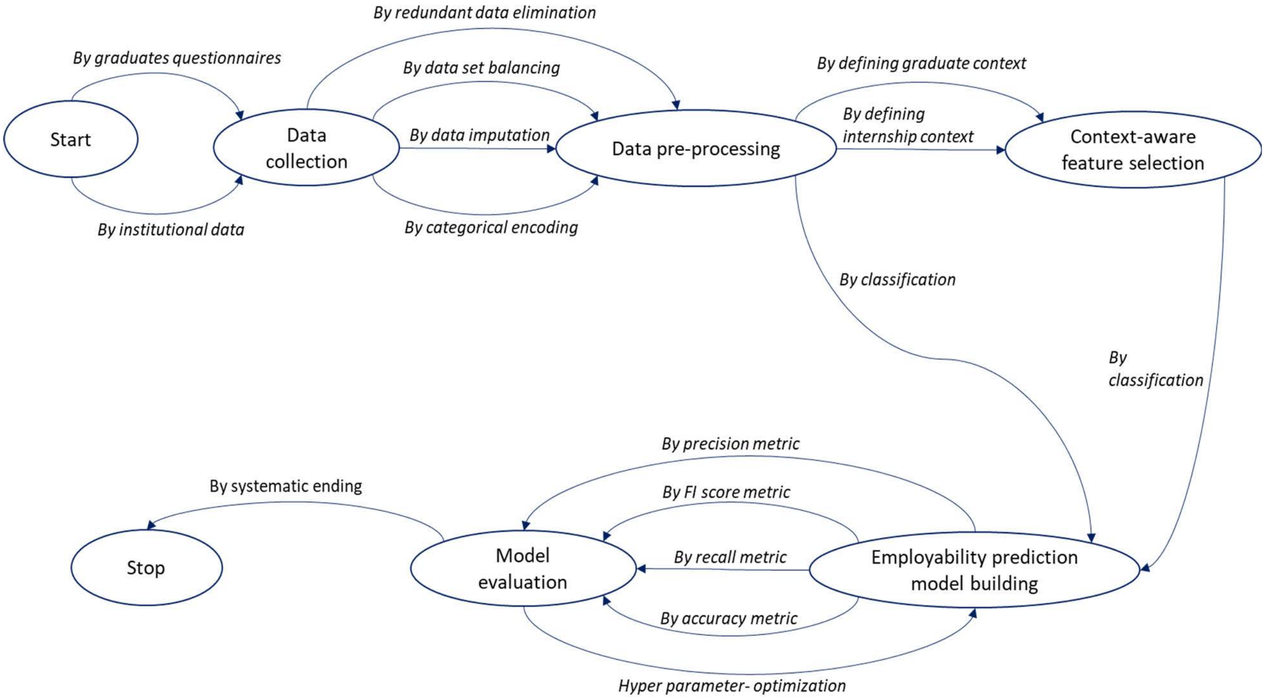

FIGURE 2. **Context-aware employability prediction process.**

concentrate on representing only the tasks. We assume that it is important to propose a ‘‘purposeful’’ ML process. We suggest extending the ‘what to do’ view and capturing the ‘Whys’. In this work, we use the intentional formalism MAP [30] to describe the main context-aware prediction process. This choice is motivated by:

- The need to introduce flexibility to achieve the different steps of the prediction process. For example, once data are pre-processed, two possibilities are given: context-aware feature selection, or employability prediction model building. Indeed, if the model has not satisfying evaluation values, the data analyst can go backward and build the ML model by using hyper parameter optimization.

- The distinction between what to achieve (i.e., the goal) and the way to achieve it (i.e., the strategy) to show the different ways allowed to move from one step to another during a standard ML prediction process. For example, we show that going from ‘‘Data collection’’ step to ‘‘Data pre-processing’’ step, many techniques are generally allowed, namely, ‘by eliminating redundant data’, ‘by data imputation’, ‘by dataset balancing’, etc. The same idea is in moving from building the model to evaluating it, where the set of evaluating metrics available for the data analyst are highlighted (i.e., precision metric, recall metric, accuracy metric, etc.).

A Map is a goal-driven navigational structure that supports the dynamic selection of the intention to be achieved next and the appropriate strategy to achieve this intention according to the current situation. A Map is a process model expressed in a goal-driven perspective and based on a non-deterministic ordering of goals and strategies. It is a decision-driven navigational structure supporting context-awareness where processes are represented in a labeled directed graph composed of intentions as nodes and strategies as edges [30]. In our work, an intention can be achieved by the performance of a ML step.

The proposed roadmap shows which ML steps can follow which ones. A Strategy is a way to achieve a ML step. A strategy S_{ij} between the couple of steps ST_i and ST_j represents the way STj can be achieved once ST_i has been satisfied. A Section is a triplet $< STi, ST_j, S_{ij} >$ and represents a way to achieve the target ML step ST_j (intention) from the source ML step ST_i (intention) following the strategy S_{ij}. Our methodology is illustrated in Fig. 2. MAP strategies are from (1): data collection techniques and sources (from the main actors involved in the internship procedure), (2) data pre-processing techniques in data mining, (3) training models, and (4) ML performance evaluation metrics. Each step is detailed in the following sections.

A. DATA COLLECTION

The data used in this paper were based on three academic years ranging from 2019 to 2021. Data were collected by distributing an online survey to graduates of the Information Systems (IS) department at the College of Computer and Information Sciences (CCIS) in Princess Nourah bint Abdulrahman University (PNU).

The survey included a total of 37 questions divided over three sections: (i) student information, (ii) internship

TABLE 5. Features description and domain.

	Feature	Description	Values domain
Student Context	GPA (General Point Grade)	The average result of all the grades is calculated by dividing the total quality points earned by the total number of credit hours for which grades were assigned.	Enum {GPA > 4.75, 4.50 < GPA < 4.75, 4 < GPA < 4.5, 3.50 < GPA < 4, 3 < GPA < 3.5, 2.50 < GPA < 3, 1 < GPA < 2}
	LinkedIn (LinkedIn account)	This feature indicates whether the student has a LinkedIn account.	Boolean {Yes, No}
	PCertificate (Professional certificates)	Certificates that prove that the person has the knowledge, experience and skills to perform a specific job.	Boolean {Yes, No}
	Co-curricular (Number of co-curricular activities)	An official document approved by the University, which monitors the life skills (personal and professional) of the student.	Integer
Internship Context	IntGrade (Internship grade)	The grade a student earned after completing the internship program.	Enum {A+, A, B+, B, C+, C, D+, D, F}
	IntFields (Internship fields)	The main fields of the internship.	Enum {Technical support, Project management, Software engineering, Business process management, Information systems management, Marketing, Governance, Database management, Data sciences, Ux design, Data analytics, Application development, Security, Data engineering, Networking, Other}
	IntDurationM (Internship duration moths)	The duration of the internship (in 'months').	Enum{1, 2, 3, 4, 5, 6, other}
	IntDurationH (Number of internship hours)	The number of training hours.	Enum {120 hrs, 120-200 hrs, >= 200 hrs}
	IntType (Internship type)	Is the internship full-time or part-time?	Enum {Full-time,Part-time}
	IntDays (Number of days per week)	The number of internship days per week.	Enum {2, 3, 4, 5}
	IntMethod (Internship method)	Is the internship online or face-to-face?	Enum {Online, face-to-face}
	IntCertRotation Internship certificate and rotation	This feature indicates whether the internship organization provides a certificate and a rotation in various departments.	Set {Certificate, Rotation, Both}
	IntSatisfaction (Internship satisfaction)	The student's satisfaction degree about internship.	Enum {1, 2, 3, 4, 5} 1: Not satisfied at all, 5: Very satisfied.
	OrgType (Internship organization type)	The domain of the organization providing the internship.	Enum {Education, Health, IT services firms, Finance, Communication, other}
	OrgSector (Internship organization sector)	Is the organization private or government?	Enum {Government, Private}
	OrgJobOffer (Internship Organization job offer)	Has the trainee received a job offer from the organization?	Boolean {Yes, No}
	OrgRecuitement (Internship Organization recruitment).	Is the student recruited by the organization offering the internship?	Boolean {Yes, No}
Output	Emp_Status: Employment status.	The student' employment status after graduation.	Enum{Employed, Unemployed, Continuing studies, Training}.

information, and (iii) employment information. After the survey was designed, 50 students were invited to fill in a survey. No significant issues were found, and an official questionnaire was eventually produced and distributed to the students.

In this study, two categories of features are distinguished: student-related features and internship-related features, as described in Table 5.

The student related features include *GPA*, *LinkedIn* (LinkedIn account), *PCertificate* (Professional Certificates), *GraduationYS* (Graduation year and semester) and Co-curricular (Number of co-curricular activities).

The internship related features include: *IntGrade* (Internship grade), *IntFields* (Internship fields), *IntDurationM* (Internship duration in months), *IntDurationH* (Number of training hours), *IntType* (Internship type), *IntDays* (Number of days per week), *IntMethod* (Internship method), *IntCertRotation* (Internship certificate and rotation), *IntSatisfaction* (Internship satisfaction), The organization related features include: *OrgType* (Internship organization type),

TABLE 6. Feature encoding.

Numerical features	Categorical features	
	One-hot-encoding	Label encoding
LinkedIn	IntGrade	IntDurationH
PCertificate	IntCertRotation	IntMethod
Co-curricular	IntFields	OrgSector
IntDurationM	OrgType	Emp_Status
IntDays		IntType
IntSatisfaction		GPA
OrgJobOffer		
OrgRecuitement		

OrgSector (Internship organization sector), *OrgJobOffer* (Internship Organization job offer) and *OrgRecuitement* (Internship Organization recruitment).

The output feature of the prediction process is *Emp_Status*, which represents the employment status after graduation.

B. DATA PRE-PROCESSING

By performing data exploration, we find that CCIS Employability dataset contains missing values for some respondent records. We handle this concern by applying the KNN imputation technique. Moreover, the dataset contains 18 features, 8 of which are numerical features and the remaining 10 are categorical features. However, ML models require all input and output features to be numeric. Therefore, the categorical features must be converted into numerical variables to fit into ML algorithms. This conversion or transformation is known as 'categorical encoding', and can severely limit the power of prediction models. Thus, such categorical features need to be meaningfully encoded better modeling and understanding.

The two most popular techniques are one-hot encoding and ordinal encoding [60]. One-hot encoding represents a categorical variable as a group of binary variables, where each binary variable represents one category. Ordinal encoding (also known as integer encoding, or label encoding) aims to transfer categorical variables to integer labels. The selection of encoding approaches depends on the type of categorical variable. In fact, there are two types of categorical variables: nominal variables having no intrinsic ordering to its categories and ordinal variables having a clear ordering. Use of ordinal encoding is recommended for categorical variables that have a natural rank ordering, while using one hot encoding for categorical variables that have no relationship between values.

To encode the categorical values, this study applies the two encoding techniques as described in Table 6.

It is worth mentioning that applying a one-hot encoding approach leads to increasing the dimensionality. For this reason, we choose to perform label encoding for *IntType, IntMethod* and *OrgSector* rather than one-hot-encoding since each of these two features has only two states. Therefore, we obtained 52 features after encoding. Furthermore, we found that our dataset is unbalanced. As illustrated in Fig. 3, we note that for the output feature "EmpStatus", class 2 is considered as the majority class, while class 3 is considered as the minority class. Typically, conventional classifiers are biased towards the majority class, leading to classification errors. To address this issue, we apply Synthetic Minority Oversampling Technique (SMOTE) [61], which aims to oversample the minority classes by synthesizing new samples from the existing samples.

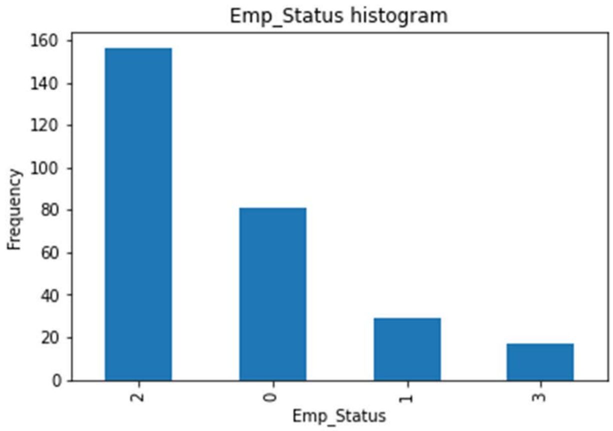

FIGURE 3. Sample distribution of the Emp_Status feature before SMOTE.

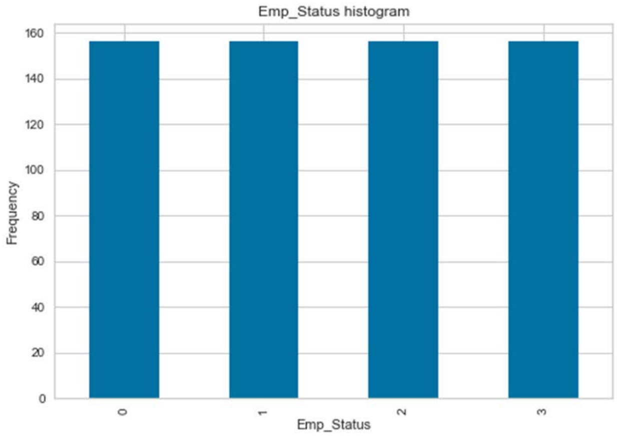

FIGURE 4. Sample distribution of the Emp_Status feature after SMOTE.

Fig.4 displays the distribution of samples for "EmpStatus" after applying SMOTE, where each class contains 156 samples. As a result, we obtained 624 total records instead of 283 initial records by generating new synthesized records.

C. CONTEXT-AWARE FEATURE SELECTION

Performing categorical encoding yields a curse of dimensionality; this rises the difficulty of designing algorithms in high dimensions and often having a running time exponential in the dimensions. Thus, it is recommended to perform feature selection to reduce the feature space and enhance the efficiency and accuracy of ML algorithms by selecting a small subset from features. This process helps identify more discriminating features and remove redundant and irrelevant features. The feature that carries no information about the various classes is said to be an irrelevant feature. If the feature has high correlation with other features and decreases the accuracy, it is said to be a redundant feature.

However, employability prediction, in real life, is the major concern of each student regardless of his or her study level. Additionally, having a specific internship condition may impact the employability of the student. For this reason, we can distinguish different context situations that would have strong correlations with the employability situation. The following examples explain the importance of context modeling. The first example is about a student who has not yet performed an internship and wants to predict his/her employability according to his/her study level. The second example is about a student who wants to know the implication of selecting a specific internship situation on his/her employability.

In this study, we opt for dealing with the two issues mentioned above by proposing a new context-aware feature selection method that helps students customize their internship choice, interactively through the prediction system, for the purpose of promoting their employability. To this end, we consider the context as the set of related factors that impact the student employability. The context is structured using the concept of a Context entity, which represents a component of the contextual information. A context entity is represented using a set of context attributes. A context attribute is an atomic measurable characteristic that makes contextual information explicit. The context entities considered in this work are Student and Internship.

- The student context is characterized by the following attributes: General Point Grade, LinkedIn account, Professional certificates, Graduation year and semester, Major and Number of co-curricular activities.

- The internship context is characterized by the following attributes: Internship grade, Internship fields, Internship duration in months, Number of training hours, Internship type, Number of days of internship per week, Internship method, Internship certificate, Rotation, Student's internship satisfaction, Internship organization type, Internship organization sector, the fact that the organization has made a job offer to the student and/or has recruited the student.

To formalize the contextual knowledge relevant to our research, we introduce a context model that aims at structuring the contextual information. The context model is based on the following main concepts: Context entity and Context attribute. The proposed context model is illustrated in Fig. 5 using the UML class diagram.

D. BUILDING EMPLOYABILITY PREDICTION MODEL

It is worth saying that building a highly reliable prediction model that has the ability to determine the employability status based on an effective feature set, is very challenging. Fig. 6 represents the prediction model selection process.

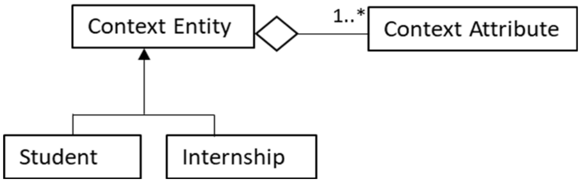

FIGURE 5. Context model for employability prediction.

Given a context C in {C_s, Ci} where:

- C_s is the student context
- C_i is the internship context,
- $FC_s = \{fs_1, \ldots fs_k\}$ is the set of features related to Cs where fs1 is the first feature in FC_s, k is the total number of features related to FC_s, and FC_s is a part from F the set of all n features
- $FC_i = \{fi_1, \ldots fi_p\}$ is the set of features related to C_i where fi_1 is the first feature in FC_i and fi_p is the p^{th} feature in FC_i, p is the total number of features related to FC_i, and FC_i is the set of all n features from F.

Our objective is to select the most suitable prediction model that achieves the best performance to our target output: employment status (Emp_Status). This output is a categorical feature that contains 4 classes. For this reason, our task is multiclass classification. Therefore, we use Three Boosting classifiers - XGBoost, LightGBM and CatBoost [62], [63] - to evaluate their performance in predicting employment status. The term 'Boosting' refers to an ensemble learning method that combines a set of weak learners into a strong learner to minimize training errors.

The XGBoost model is a scalable DT-based ensemble algorithm that is based upon the concept of gradient tree boosting. It is a tree ensemble approach where the trees are added sequentially and each tree learns from its predecessors where they aim to minimize the errors of the previous tree. The trees are provided in parallel tree boosting to solve tasks in a fast and accurate way. Since its introduction, it has been improving its efficiency and power for solving a broad range of classification problems.

LightGBM algorithm is a boosting algorithm that depends on DT algorithms. It divides the tree leaf wise with the best fit though other boosting algorithms split the tree profundity wise or level wise as opposed to leaf-wise". LightGBM involves mainly Gradient Boosting Decision Tree (GBDT), which is an algorithm that involves a boosting technique. In the literature, the LightGBM model was considered to have the highest accuracy.

The CatBoost model places more emphasis on, as the name suggests, categorical features. CatBoost combats the issue of exponential feature combination growth by using a greedy method at every new split of the current tree. CatBoost is also able to handle scenarios where the number of categories would be too large for other GBDT models. CatBoost excels in dealing with non-numeric data, as it can convert categorical values to integers in a way it deems to be the optimal approach.

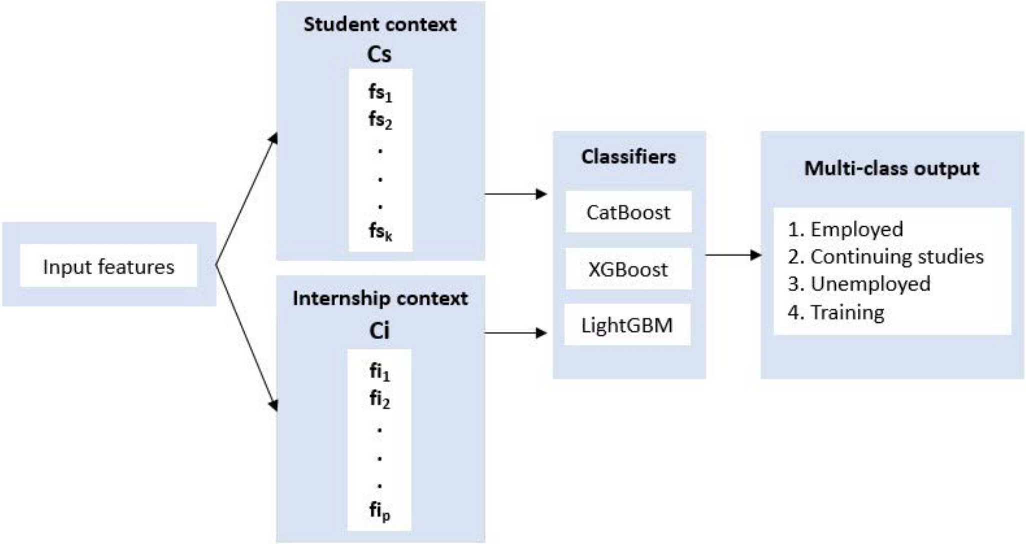

FIGURE 6. Training model selection process.

CatBoost was developed to handle categorical features; performances may differ significantly depending on the types of features implemented.

E. MODEL EVALUATION

We applied Repeated K-fold cross-validation to estimate the performance of each boosting model, which involves simply repeating the cross-validation procedure multiple times and reporting the mean result across all folds from all runs.

The number k denotes the number of folds into which our dataset is divided. In this paper, we consider k = 10 and the process is referred to as 10-fold cross-validation.

Four evaluation indicators are used to evaluate and compare the quality of the models: accuracy, precision, F-score and recall. The accuracy is the percent of correct predictions for the test data; and is calculated by dividing the number of correct predictions by the number of total predictions. The precision is the ratio of correctly predicted positive examples divided by the total number of positive examples that were predicted. The recall is the ratio of the number of relevant predictions retrieved to the total number of relevant examples in the test dataset. The F-score is defined as the harmonic mean of the precision and recall of the model. It is designed to be a useful metric when classifying unbalanced classes.

IV. EXPERIMENTS AND RESULTS

The implementation of our proposed process was performed using the Scikit-Learn and PyCaret libraries in Python 3.7. We built our experimentation workflow according to the research questions posed in the introduction.

A. COMPARISON OF BOOSTING MODELS VS. ORDINARY MODELS

In this experiment, we compared the selected boosting models using all features, in terms of accuracy, precision, recall and F-score. Table 7 displays the results obtained. We note that CatBoost, RF, XGBoost, Extra Trees Classifier and LightGBM outperform the other ML algorithms for all selected performance metrics, while ordinal classifiers such as SVM, KNN and NB achieve worse scores. Therefore, the obtained outcomes confirm our theoretical findings about the power of boosting algorithms in terms of improving model predictions for learning algorithms by sequentially correcting the weak learners from their predecessors to be converted into strong learners.

TABLE 7. Performance evaluation of different ML models using all features.

Model	Accuracy	Recall	Precision	F-score
LightGBM	0.7753	0.7739	0.7894	0.7678
CatBoost	0.7591	0.7586	0.7734	0.7564
XGBoost	0.7408	0.7398	0.7494	0.7343
DT	0.6672	0.6661	0.6843	0.6585
LR	0.6101	0.6103	0.6260	0.6076
LDA	0.6032	0.6041	0.6074	0.5945
KNN	0.5892	0.5877	0.6035	0.5702
AdaBoost	0.5686	0.5700	0.5803	0.5585
SVM	0.4701	0.4695	0.5366	0.4606
NB	0.4401	0.4422	0.5033	0.3929

Fig. 7, Fig. 8 and Fig. 9 illustrate the Class Prediction Errors of the selected boosting models. The purpose of this experiment is to compare their performance in terms of the number of correct and incorrect predictions.

By examining these figures, we note that these boosting algorithms have similar results of correct predictions for classes 1 and 3, which are ‘training’ and ‘continuing studies’. Additionally, we note that the prediction of class 2

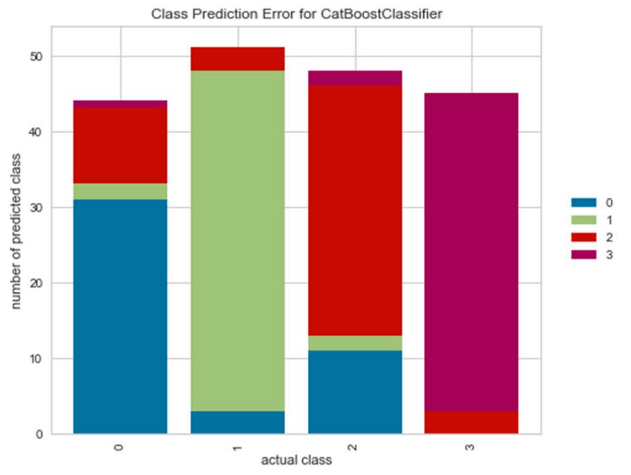

FIGURE 7. Class prediction error for the CatBoost model.

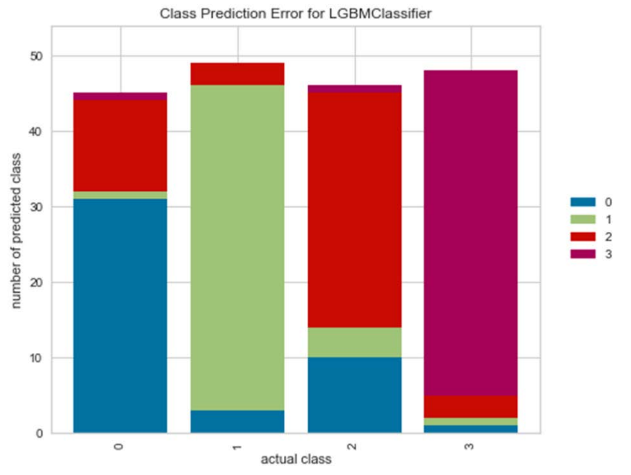

FIGURE 8. Class prediction error for the LightGBM model.

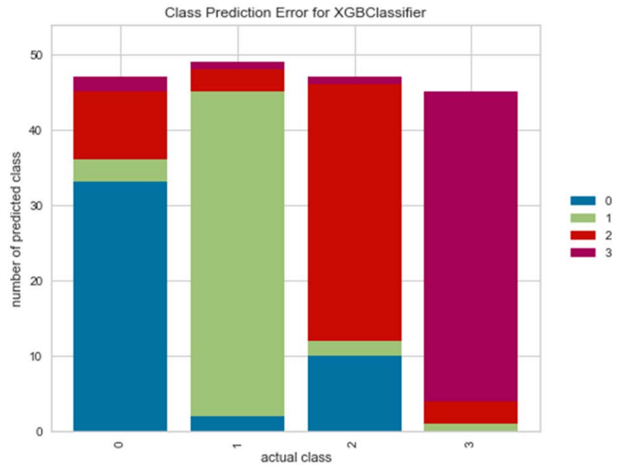

FIGURE 9. Class prediction error for the XGBoost model.

'employed' is confused with Class 0 'unemployed' for all models.

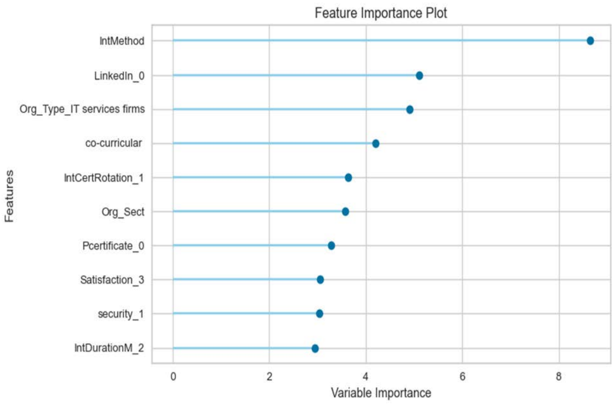

FIGURE 10. Feature selection using the CatBoost method.

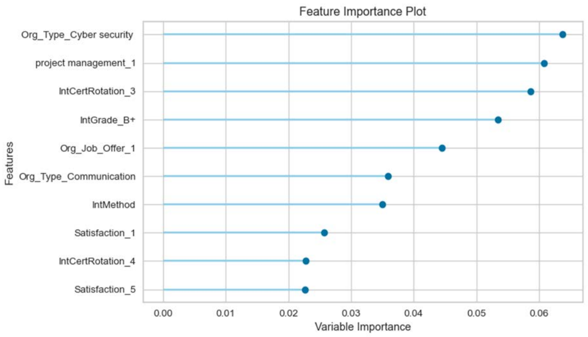

FIGURE 11. Feature selection using the XGBoost method.

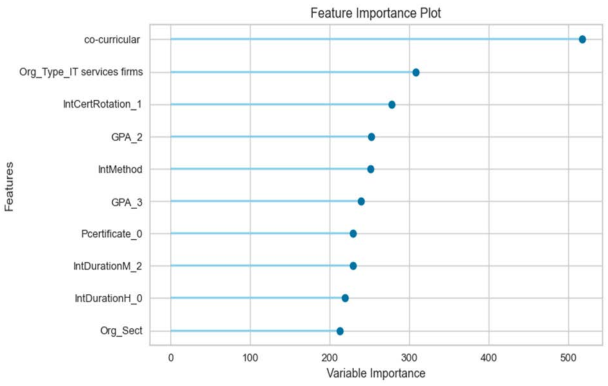

FIGURE 12. Feature selection using the LightGBM method.

B. ANALYSIS OF TOP SENSITIVE FEATURES FOR SELECTED MODELS

In this experiment, we aim to identify the most important features that help improve the employment status prediction task.

Fig. 10, Fig. 11 and Fig. 12 display the top 10-ranked features using the Feature Importance method applied to the CatBoost, XgBoost and LightGBM classifiers respectively. The feature importance method assigns weights to features, and then ranks them based on their weights.

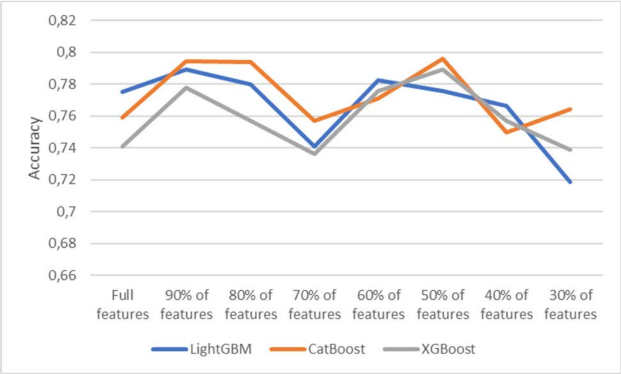

FIGURE 13. Accuracy evaluation of different classifiers after applying feature selection at different thresholds.

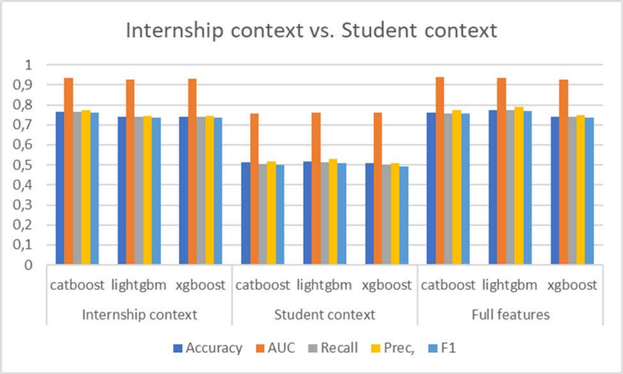

FIGURE 14. Evaluation of internship context and student context impact on predicting student employability.

From Fig. 10, we note that the 'intMethod' feature has a great impact on the employability status prediction using the CatBoost model. In Fig. 11, 'org_type_Cyber security' and 'project management_1' features have the highest importance score; this means that performing the internship on cybersecurity organization will influence the student employability; moreover, the internship on the project management field plays an important role in the employability of students. In Fig. 12, the 'co-curricular' feature has the greatest importance value in the LightGBM model. Therefore, the co-curricular activities of students have a great influence on their employability.

To evaluate the impact of feature selection on the performance of the prediction models, we compare the accuracy metric of each model with vs. without the use of feature selection. The obtained results are displayed in Fig. 13.

According to Fig. 13, we observe that feature selection can enhance the accuracy of the prediction models. The improvement of accuracy score for all models is seen when selecting 90% of the features compared to all features. Moreover, when selecting the top 50% of features, the accuracy increased from 0.7591 to 0.7959 for the CatBoost classifier and from 0.7408 to 0.7891 for the XGBoost classifier. However, when selecting 60% of features, the accuracy increased from 0.7753 to 0.7825 for the LightGBM classifier. Accordingly, we can conclude that by applying an automatic feature selection method, we can obtain better performance while reducing the complexity space by eliminating irrelevant features. In the next experiments, we aim to study the impact of semantically selecting the features according to their contexts, mainly the internship context and the student context.

C. EVALUATION OF INTERNSHIP CONTEXT AND STUDENT CONTEXT IMPACTS ON EMPLOYABILITY PREDICTION TASKS

As previously explained in Section III-B, the internship context is constituted by features related to internship, while the student context consists of student features. These features are semantically correlated since they belong to the same context. In this experiment, we evaluate the internship context and student context impacts on improving employability status prediction models. The obtained results are displayed in Fig. 14.

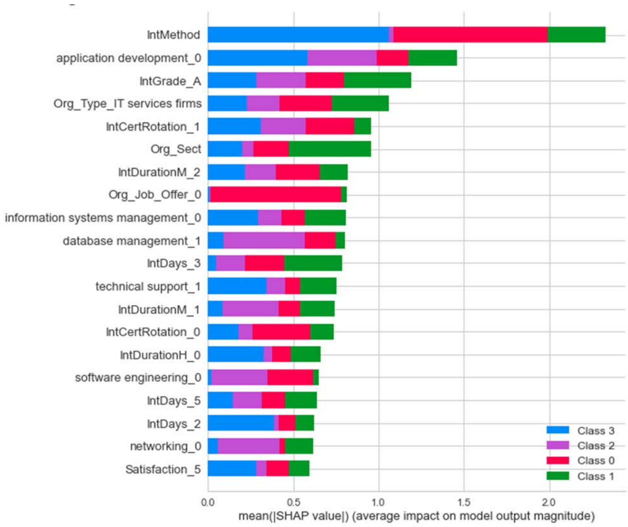

FIGURE 15. SHAP analysis of the XGBoost model.

According to Fig. 14, the internship context provides better results than the student context. In fact, by having only internship context, we can predict employment status with 76% accuracy. Meanwhile, the student context performs the worst by achieving only 51% accuracy on average, indicating that the features defining the student context are insufficient to predict the employability status. Thus, the student context should be further investigated by adding more relevant features that can help the model achieve better results.

Furthermore, it is clear that internship context is the key factor determining the future of graduates in terms of employability. To better understand the role of each internship feature in determining the employability status, we perform the SHAP (SHapley Additive exPlanations) model [64] for each model as presented in Fig. 15, Fig. 16 and Fig. 17.

Fig. 15, Fig. 16 and Fig. 17 display the SHAP summary plots, for the XGBoost, CatBoost and LightGBM

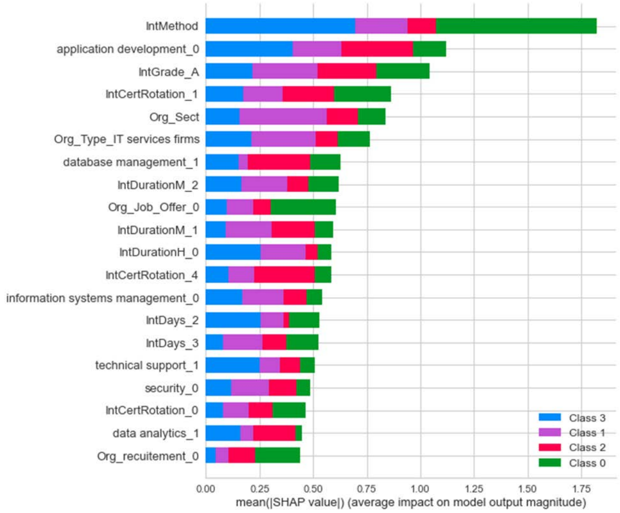

FIGURE 16. SHAP analysis of the CatBoost model.

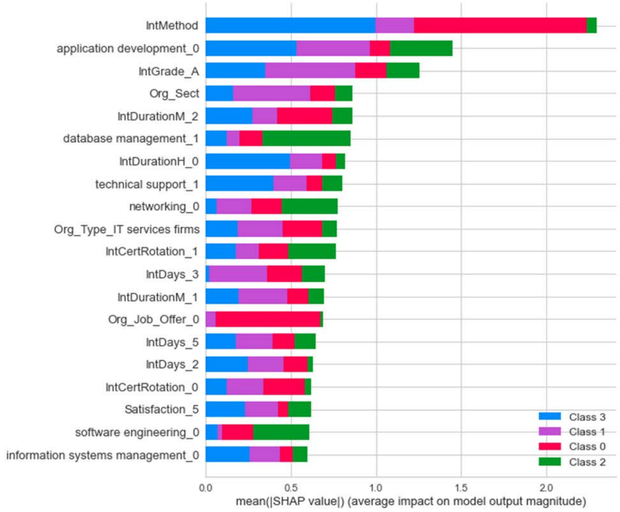

FIGURE 17. SHAP analysis of the LightGBM model.

models, respectively. These figures illustrate the contribution of each feature in predicting each class. It is clear that, for the three models, the 'Intmethod' feature plays the most important role in predicting classes 3 and 0, which are the 'training' and 'employed' classes, respectively, while 'Org_Job_offers_0' exclusively determines class 0 for the LightGBM and XGBoost models.

Additionaly, we observe that 'IntGrade_A' and 'Org_Sect' features contribute to predicting Class 1, which is 'Continuing Studies'. Furthermore, we observe that the 'database management_1' feature has a strong correlation with the class 2 'Unemployed'. Unfortunately, we cannot estimate whether training on database management will strongly affirm that the student will be unemployed or inversely.

V. DISCUSSION

For our first research question, "What are the best ML algorithms for employability prediction?", we found that XGBoost, CatBoost and LightGBM have the best performance values, as they are different implementations of gradient boosting algorithms. In addition, most of gradient boosting algorithms provide support for handling categorical features. This fact contrasts the retained studies in the literature review, except Laddha *et al.*, which used AdaBoosting and other algorithms for employability prediction.

Referring to the second raised research question "What are the top sensitive features that impact employability prediction?", we found that the most sensitive features for the three ML algorithms used fit into the Internship category feature.

Moreover, for CatBoost and LightGBM the remaining sensitive features fit into: (i) co-curricular activities, which supports the findings of Dubey and Mani [49] and Othman et al [20], and (ii) soft skills, as in [20]. However, hard skills are considered sensitive features only for LightGBM, as in the research results of Hugo [19].

Concerning the third question that investigates "Do student context and internship context influence the employability status prediction"? First, the results obtained show that using context with ML for student employability prediction is feasible and effective. Indeed, our findings reveal that the internship context has more influence on employability status prediction compared to the student context.

For students of information systems in computer sciences colleges, the determinant attributes to predict their employability status are mainly related to the internship method (face to face or online), the grade they obtained in the internship subject, whether the student has obtained a certificate after achieving the internship, etc. This determinant is in line with the major requirements for employment in the government or private Saudi workforce, which give high importance to student internship experience, as reported in [2], [33], [36], [37].

VI. CONCLUSION

In this study, we presented a new approach to predict student employability based on context and using Gradient Boosting Models. We introduced a context-aware employability prediction process for structuring the employability prediction steps. We proposed a context model for employability prediction and defined a context-based ML method for predicting employability status of students. We also identified the most predictive features impacting employability. Experiments were conducted using three gradient boosting classifiers: XGBoost, CatBoost and LightGBM. The results of the experiments show that the internship context plays the most important role in predicting the employability status in comparison to the student context. In addition, the experiments demonstrate promising results achieved by applying Boosting models, mainly the LightGBM model which performed the best. In conclusion, internship features are the most sensitive features among inputs that affect the employability status of students.

The valuable contributions of this work can be helpful for both students, supervisors, internship authorities and coordinators, instructors, and programme directors, in many ways: (i) anticipate and improve curriculum (e.g. credit hours), for forthcoming years; (ii) re-view internship process and requirements (e.g. increase training hours, rigorously select internship organizations, and favor those that offer certificates, favor rotation programs); (iii) discover student strengths and weaknesses that they need to overcome for employment; and (iv) identify students at risk of unemployment at an early stage.

This study presents the limitation of predicting the employability status through internship of students in computer and information sciences fields and lacks other attributes for students enrolling in other disciplines. Another limitation is that this study does not consider the internship context for students with special needs who face lower employment rates. Additional attributes can affect the prediction of employability status for graduates with disabilities, such as language skills and type of disability.

We think that there are still improvement opportunities for predicting student employability through the internship context. The student context can be enriched with additional relevant student features, such as socio-demographic characteristics and grades obtained in enrolled curriculum subjects, which will increase the student context influence on employability prediction even to the internship context. Data from site internship evaluation are also interesting to include in internship context attributes, such as collaboration, commitment to work and leadership. Ultimately, we recommend extending this work by predicting job search duration, salary range and employment field of students, through the internship context.

ACKNOWLEDGMENT

The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number PNU-DRI-RI-20-003.

REFERENCES

[1] A. Zopiatis and J. Pribic, "College students' dining expectations in Cyprus," *Brit. Food J.*, vol. 109, no. 10, pp. 765–776, Oct. 2007, doi: 10.1108/00070700710821313.

[2] D. A. Sapp and Q. Zhang, "Trends in industry supervisors' feedback on business communication internships," *Bus. Commun. Quart.*, vol. 72, no. 3, pp. 274–288, Sep. 2009, doi: 10.1177/1080569909336450.

[3] S. B. Knouse and G. Fontenot, "Benefits of the business college internship: A research review," *J. Employment Counseling*, vol. 45, no. 2, pp. 61–66, Jun. 2008, doi: 10.1002/j.2161-1920.2008.tb00045.x.

[4] P. Stansbie, R. Nash, and S. Chang, "Linking internships and classroom learning: A case study examination of hospitality and tourism management students," *J. Hospitality, Leisure, Sport Tourism Educ.*, vol. 19, pp. 19–29, Nov. 2016, doi: 10.1016/j.jhlste.2016.07.001.

[5] L. Ruhanen, R. Robinson, and N. Breakey, "A foreign assignment: Internships and international students," *J. Hospitality Tourism Manage.*, vol. 20, pp. 1–4, Jan. 2013, doi: 10.1016/j.jhtm.2013.05.005.

[6] L. Ruhanen, R. Robinson, and N. Breakey, "A tourism immersion internship: Student expectations, experiences and satisfaction," *J. Hospitality, Leisure, Sport Tourism Educ.*, vol. 13, pp. 60–69, Jul. 2013, doi: 10.1016/j.jhlste.2013.02.001.

[7] T. N. Garavan and C. Murphy, "The co-operative education process and organisational socialisation: A qualitative study of student perceptions of its effectiveness," *Educ. Training*, vol. 43, no. 6, pp. 281–302, Sep. 2001, doi: 10.1108/EUM0000000005750.

[8] H. Yang, C. Cheung, and H. Song, "Enhancing the learning and employability of hospitality graduates in China," *J. Hospitality, Leisure, Sport Tourism Educ.*, vol. 19, pp. 85–96, Nov. 2016, doi: 10.1016/j.jhlste.2016.08.004.

[9] E. Ishengoma and T. I. Vaaland, "Can university-industry linkages stimulate student employability?" *Educ. Training*, vol. 58, no. 1, pp. 18–44, Jan. 2016, doi: 10.1108/ET-11-2014-0137.

[10] E. Qenani, N. MacDougall, and C. Sexton, "An empirical study of self-perceived employability: Improving the prospects for student employment success in an uncertain environment," *Act. Learn. Higher Educ.*, vol. 15, no. 3, pp. 199–213, Nov. 2014, doi: 10.1177/1469787414544875.

[11] A. Ring, A. Dickinger, and K. Wöber, "Designing the ideal undergraduate program in tourism: Expectations from industry and educators," *J. Travel Res.*, vol. 48, no. 1, pp. 106–121, Aug. 2009, doi: 10.1177/0047287508328789.

[12] T. Lam and L. Ching, "An exploratory study of an internship program: The case of Hong Kong students," *Int. J. Hospitality Manage.*, vol. 26, no. 2, pp. 336–351, Jun. 2007, doi: 10.1016/j.ijhm.2006.01.001.

[13] N. Mezhoudi, R. Alghamdi, R. Aljunaid, G. Krichna, and D. Düştegör, "Employability prediction: A survey of current approaches, research challenges and applications," *J. Ambient Intell. Humanized Comput.*, pp. 1–17, Jun. 2021, doi: 10.1007/s12652-021-03276-9.

[14] H. O'Connor and M. Bodicoat, "Exploitation or opportunity? Student perceptions of internships in enhancing employability skills," *Brit. J. Sociol. Educ.*, vol. 38, pp. 1–12, Dec. 2015, doi: 10.1080/01425692.2015.1113855.

[15] V. Gamboa, M. P. Paixão, and S. N. de Jesus, "Internship quality predicts career exploration of high school students," *J. Vocational Behav.*, vol. 83, no. 1, pp. 78–87, Aug. 2013, doi: 10.1016/j.jvb.2013.02.009.

[16] L. H. N. Fong, H. A. Lee, C. Luk, and R. Law, "How do hotel and tourism students select internship employers? A segmentation approach," *J. Hospitality, Leisure, Sport Tourism Educ.*, vol. 15, pp. 68–79, Jul. 2014.

[17] J. Gault, J. Redington, and T. Schlager, "Undergraduate business internships and career success: Are they related?" *J. Marketing Educ.*, vol. 22, no. 1, pp. 45–53, Apr. 2000, doi: 10.1177/0273475300221006.

[18] J. Gault, E. Leach, and M. Duey, "Effects of business internships on job marketability: The employers' perspective," *Educ. Training*, vol. 52, no. 1, pp. 76–88, Feb. 2010, doi: 10.1108/00400911011017690.

[19] L. S. Hugo, "Predicting employment through machine learning," *NACE J.*, May 2019. [Online]. Available: https://www.naceweb.org/career-development/trends-and-predictions/predicting-employment-through-machine-learning/

[20] Z. Othman, S. W. Shan, I. Yusoff, and C. P. Kee, "Classification techniques for predicting graduate employability," *Int. J. Adv. Sci., Eng. Inf. Technol.*, vol. 8, nos. 4–2, p. 1712, Sep. 2018, doi: 10.18517/ijaseit.8.4-2.6832.

[21] J. M. Nunley, A. Pugh, N. Romero, and R. A. Seals, "College major, internship experience, and employment opportunities: Estimates from a résumé audit," *Labour Econ.*, vol. 38, pp. 37–46, Jan. 2016.

[22] A. Ksibi, A. B. Ammar, and C. B. Amar, "Enhanced context-based query-to-concept mapping in social image retrieval," in *Proc. Int. Workshop Content-Based Multimedia Indexing (CBMI)*, Jun. 2013, pp. 85–89, doi: 10.1109/CBMI.2013.6576559.

[23] R. Fakhfakh, A. Ksibi, A. B. Ammar, and C. B. Amar, "Enhancing query interpretation by combining textual and visual analyses," in *Proc. Int. Conf. Adv. Logistics Transp.*, May 2013, pp. 170–175, doi: 10.1109/ICAdLT.2013.6568454.

[24] A. Dey, "Understanding and using context," *Pers. Ubiquitous Comput.*, vol. 5, pp. 4–7, Feb. 2001, doi: 10.1007/s007790170019.

[25] O. Saidani and S. Nurcan, "Context-awareness for adequate business process modelling," in *Proc. 3rd Int. Conf. Res. Challenges Inf. Sci.*, Apr. 2009, pp. 177–186, doi: 10.1109/RCIS.2009.5089281.

[26] L. Jamel, O. Saidani, and S. Nurcan, "Flexibility in business process modeling to deal with context-awareness in business process reengineering projects," in *Proc. 20th Enterprise, Bus.-Process Inf. Syst. Modeling*, Tallinn, Estonia, Jun. 2018, pp. 35–48.

[27] O. Saidani, C. Rolland, and S. Nurcan, "Towards a generic context model for BPM," in *Proc. 48th Hawaii Int. Conf. Syst. Sci.*, Jan. 2015, pp. 4120–4129, doi: 10.1109/HICSS.2015.494.

[28] S. Najar, O. Saidani, M. K. Pinheiro, C. Souveyet, and S. Nurcan, "Semantic representation of context models: A framework for analyzing and understanding," in *Proc. 1st Workshop Context, Inf. Ontologies.*, Jun. 2009, pp. 1–10, doi: 10.1145/1552262.1552268.

[29] N. Nascimento, P. Alencar, C. Lucena, and D. Cowan, "A context-aware machine learning-based approach," in *Proc. 28th Annu. Int. Conf. Comput. Sci. Softw. Eng. (CASCON)*, Oct. 2018, pp. 40–47.

[30] C. Rolland and S. Nurcan, ''Business process lines to deal with the variability,'' in *Proc. 43rd Hawaii Int. Conf. Syst. Sci.*, 2010, pp. 1–10.

[31] O. Saidani and S. Nurcan, ''Multi-level delegation for flexible business process modeling,'' in *Proc. Int. Conf. Inf. Resour. Manage. Assoc., Bus. Process. Manage. Track*, Vancouver, BC, Canada, May 2007, pp. 1–11.

[32] I. B. Fraj, Y. B. Hlaoui, and L. BenAyed, ''A reactive system for specifying and running flexible cloud service business processes based on machine learning,'' in *Proc. IEEE 45th Annu. Comput., Softw., Appl. Conf. (COMPSAC)*, Jul. 2021, pp. 1483–1489, doi: 10.1109/COMPSAC51774.2021.00220.

[33] A. Zopiatis and A. L. Theocharous, ''Revisiting hospitality internship practices: A holistic investigation,'' *J. Hospitality, Leisure, Sport Tourism Educ.*, vol. 13, pp. 33–46, Jul. 2013.

[34] H.-B. Kim and E. J. Park, ''The role of social experience in undergraduates' career perceptions through internships,'' *J. Hospitality, Leisure, Sport Tourism Educ.*, vol. 12, no. 1, pp. 70–78, Apr. 2013, doi: 10.1016/j.jhlste.2012.11.003.

[35] G. Siu, C. Cheung, and R. Law, ''Developing a conceptual framework for measuring future career intention of hotel interns,'' *J. Teaching Travel Tourism*, vol. 12, no. 2, pp. 188–215, Apr. 2012, doi: 10.1080/15313220.2012.678220.

[36] T.-L. Chen and C.-C. Shen, ''Today's intern, tomorrow's practitioner?—The influence of internship programmes on students' career development in the hospitality industry,'' *J. Hospitality, Leisure, Sport Tourism Educ.*, vol. 11, no. 1, pp. 29–40, Apr. 2012, doi: 10.1016/j.jhlste.2012.02.008.

[37] T.-L. Chen, C.-C. Shen, and M. Gosling, ''Does employability increase with internship satisfaction? Enhanced employability and internship satisfaction in a hospitality program,'' *J. Hospitality, Leisure, Sport Tourism Educ.*, vol. 22, pp. 88–99, Jun. 2018, doi: 10.1016/j.jhlste.2018.04.001.

[38] S. T. Eurico, J. A. M. da Silva, and P. O. do Valle, ''A model of graduates' satisfaction and loyalty in tourism higher education: The role of employability,'' *J. Hospitality, Leisure, Sport Tourism Educ.*, vol. 16, pp. 30–42, Jun. 2015, doi: 10.1016/j.jhlste.2014.07.002.

[39] S. Tiwari, P. Chanak, and S. K. Singh, ''A review of the machine learning algorithms for COVID-19 case analysis,'' *IEEE Trans. Artif. Intell.*, early access, Jan. 11, 2022, doi: 10.1109/TAI.2022.3142241.

[40] O. Saidani, R. R. Manoharan, A. S. Naje, R. Mishra, A. Subburaj, S. Maheswari, M. Y. Sikkandar, S. G. Sundaram, R. Rajan, and S. Sengan, ''Analysis of positive correlation in magnitude and time measurement for earthquake using electric signals,'' *Earth Sci. Informat.*, Jan. 2022 doi: 10.1007/s12145-021-00754-8.

[41] O. S. Neffati and O. Saidani, ''Semantic web service discovery approaches: A comparative study,'' *Int. J. Comput. Sci. Mobile Comput.*, vol. 9, no. 2, pp. 87–95, Feb. 2020.

[42] L. J. Menzli, S. Ayouni, and M. Elsadig, ''Quality assessment of business process models: Case of accreditation process in higher education,'' *Int. J. Civil Eng. Technol.*, vol. 10, pp. 874–884, May 2019.

[43] K. S. Bhagavan, J. Thangakumar, and D. V. Subramanian, ''Predictive analysis of student academic performance and employability chances using HLVQ algorithm,'' *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 3, pp. 3789–3797, Mar. 2021, doi: 10.1007/s12652-019-01674-8.

[44] J. K. J. Kalpana and K. Venkatalakshmi, ''Intellectual performance analysis of students by using data mining techniques,'' *Int. J. Innov. Res. Sci., Eng. Technol.*, vol. 3, pp. 1–8, Mar. 2014.

[45] C. D. Casuat, ''Predicting Students' employability using support vector machine: A SMOTE-optimized machine learning system,'' *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 5, pp. 2101–2106, May 2020, doi: 10.30534/ijeter/2020/102852020.

[46] A. Patel, S. Mascarenhas, A. Thomas, and D. Varghese, ''Student performance analysis and prediction of employable domains using machine learning,'' in *Proc. Int. Conf. Recent Adv. Comput. Techn. (IC-RACT)*, Jun. 2020. [Online]. Available:https://ssrn.com/abstract=3682499, doi: 10.2139/ssrn.3682499.

[47] C. D. Casuat and E. D. Festijo, ''Identifying the most predictive attributes among employability signals of undergraduate students,'' in *Proc. 16th IEEE Int. Colloq. Signal Process. Appl. (CSPA)*, Feb. 2020, pp. 203–206, doi: 10.1109/CSPA48992.2020.9068681.

[48] A. Bai and S. Hira, ''An intelligent hybrid deep belief network model for predicting students employability,'' *Soft Comput.*, vol. 25, no. 14, pp. 9241–9254, Jul. 2021, doi: 10.1007/s00500-021-05850-x.

[49] A. Dubey and M. Mani, ''Using machine learning to predict high school student employability—A case study,'' in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2019, pp. 604–605, doi: 10.1109/DSAA.2019.00078.

[50] L. H. Pinto and D. C. Ramalheira, ''Perceived employability of business graduates: The effect of academic performance and extracurricular activities,'' *J. Vocational Behav.*, vol. 99, pp. 165–178, Apr. 2017.

[51] B. Jantawan and C.-F. Tsai, ''The application of data mining to build classification model for predicting graduate employment,'' *Int. J. Comput. Sci. Inf. Secur.*, vol. 11, no. 10, pp. 1–7, Dec. 2013.

[52] A. Giri, M. V. V. Bhagavath, B. Pruthvi, and N. Dubey, ''A placement prediction system using k-nearest neighbors classifier,'' in *Proc. 2nd Int. Conf. Cognit. Comput. Inf. Process. (CCIP)*, Aug. 2016, pp. 1–4.

[53] E. Osmanbegovic and M. Suljic, ''Data mining approach for predicting student performance,'' *Econ. Rev., J. Econ.*, vol. 10, no. 1, pp. 3–12, May 2012.

[54] M. D. Laddha, V. T. Lokare, A. W. Kiwelekar, and L. D. Netak, ''Performance analysis of the impact of technical skills on employability,'' *Int. J. Performability Eng.*, vol. 17, no. 4, pp. 371–378, Apr. 2021. [Online]. Available: http://www.ijpe-online.com/EN/Y2021/V17/I4/371

[55] M. S. Kumar and G. P. Babu, ''Comparative study of various supervised machine learning algorithms for an early effective prediction of the employability of students,'' *J. Eng. Sci.*, vol. 10, pp. 240–251, Oct. 2019.

[56] D. J. M. Reddy, S. Regella, and S. R. Seelam, ''Recruitment prediction using machine learning,'' in *Proc. 5th Int. Conf. Comput., Commun. Secur. (ICCCS)*, Oct. 2020, pp. 1–4, doi: 10.1109/ICCCS49678.2020.9276955.

[57] K. B. Aviso, J. I. B. Janairo, R. I. G. Lucas, M. A. B. Promentilla, D. Ethelbhert, C. Yu, and R. R. Tan, ''Predicting higher education outcomes with hyperbox machine learning: What factors influence graduate employability?'' *Chem. Eng. Trans.*, vol. 81, pp. 679–684, 2020.

[58] S. Maheswari, ''A review on predicting student performance using deep learning technique,'' Tierärztliche Praxis, Tech. Rep., 2020, vol. 40.

[59] M. M. Almutairi and M. H. A. Hasanat, ''Predicting the suitability of IS students' skills for the recruitment in Saudi Arabian industry,'' in *Proc. 21st Saudi Comput. Soc. Nat. Comput. Conf. (NCC)*, Apr. 2018, pp. 1–6.

[60] F. Pargent, F. Pfisterer, J. Thomas, and B. Bischl, ''Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features,'' 2021, *arXiv:2104.00629*.

[61] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, ''Smote: Synthetic minority over-sampling technique,'' *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Dec. 2002.

[62] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, ''A comparative analysis of gradient boosting algorithms,'' *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.

[63] L. K. Smirani, H. A. Yamani, L. J. Menzli, and J. A. Boulahia, ''Using ensemble learning algorithms to predict Student failure and enabling customized educational paths,'' *Sci. Program.*, vol. 2022, pp. 1–15, Apr. 2022.

[64] S. Lundberg and S. Lee, ''A unified approach to interpreting model predictions,'' in *Proc. Adv. Neural Inf. Process. Syst.*, Dec 2017, pp. 4768–4777.

OUMAIMA SAIDANI (Member, IEEE) received the Ph.D. degree in computer science from Paris 1-Pantheon Sorbonne University, Paris, France, and the M.Sc. degree in computer science from Paris 9—Dauphine University, France.

Then, she joined the High Institute of Computer Science, Kef, Tunisia, as an Assistant Professor. She is currently an Assistant Professor with the Information Systems Department, College of Computer and Information Sciences (CCIS-IS), Princess Nourah bint Abdulrahman University (PNU), South Korea. In addition, she is also a Researcher with the RIADI Laboratory, Tunisia. Her research interests include information systems engineering, business process engineering, process mining, context-aware computing, machine learning. Findings from her research have been published in more than 20 refereed publications. She was a member of the Scientific Committee and/or an organization committee of a number of international conferences.

LEILA JAMEL MENZLI is an Assistant Professor in the Information Systems department-College of Computer and Information Sciences at Princess Nourah Bint Abdulrahman University (PNU), KSA. She is a researcher in RIADI Laboratory-Tunisia. She received a Ph.D degree in Computer Sciences and Information Systems and an engineering degree in Computer Sciences. Her research interests are: Business Process Reengineering, Process Modeling, BPM, Data Sciences, ML, Process Mining, e-learning and software engineering. She was the Program Leader of the IS program and the ABET and NCAAA accreditation committees in the CCIS- at PNU. She worked as a head of the Information Systems Security department at the Premier Ministry in Tunisia. She is the reviewer of many international journals and conferences. She was a member of scientific/steering committees of many international conferences.

AMEL KSIBI received the B.S., M.S., and Ph.D. degrees in computer engineering from the National School of Engineering of Sfax (ENIS), Sfax University, Tunisia, in 2008, 2010, and 2014, respectively. She spent three years at ENIS as a Teaching Assistant, before joining the Higher Institute of Computer Science and Multimedia Gabes (ISIMG) as a Permanent Assistant, in 2013. She joined the Computer Science Department, Umm Qura University (UQU), as an Assistant Professor, in 2014. After, she joined Princess Nourah bint Abdulrahman University, in 2018, where she is currently an Assistant Professor with the Department of Information Systems, College of Computer Sciences and Information. Her research interests include computer vision, image processing, deep learning, information retrieval, lifelogging and wellbeing, smart education, smart agriculture, and sustainable environment.

NAZIK ALTURKI received the Ph.D. degree in information systems from The University of Melbourne. She is currently an Assistant Professor with the Information Systems Department, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Her research interests include health informatics, big data, data analytics, and mining.

ALA SALEH ALLUHAIDAN works as an Assistant Professor in the Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Saudi Arabia. Her research interests include data analytics, machine learning, and IOT security. Her background is a B.Sc. degree in Computer Science from the same university she works currently. She received her M.Sc. degree in Computer Information Systems from the Grand Valley State University, MI USA, and her Ph.D. degree in Information Systems and Technology from Claremont Graduate University, CA USA.