Employability prediction: a survey of current approaches, research challenges and applications

Nesrine Mezhoudi¹ · Rawan Alghamdi¹ · Rim Aljunaid² · Gomathi Krichna¹ · Dilek Düşteğör²

Received: 19 October 2020 / Accepted: 15 April 2021 / Published online: 16 June 2021

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

Student employability is crucial for educational institutions as it is often used as a metric for their success. The job market landscape, however, more than ever dynamic, is evolving due to the globalization, automation, and recent advances in Artificial Intelligence. Identifying the significant factors affecting employability, as well as the requirements of the new job market can tremendously help all stakeholders. Knowing their weaknesses and strengths, students might better plan their career. Instructors can focus on more appropriate skill sets to meet the requirements of rapidly evolving labor markets. Program managers can anticipate and improve their curriculum to build new competencies, both for educating, training and reskilling current and future workers. All these combined efforts certainly can contribute to increasing employability. Data driven and machine learning techniques have been extensively used in various fields of educational data mining. More and more studies are investigating data mining techniques for the prediction of employability. Yet, these studies show a lot of variation, for instance, with respect to the data used, the methods adopted, or even the research questions posed. In this paper, we aim to depict a clear picture of the art, clarifying for each standard step of data mining process, the differences, and similarities of these studies, along with further suggestions. Thus, this survey provides a comprehensive roadmap, enabling the application of data mining for employability.

Keywords Data mining · Employability · Machine learning · Prediction · Review

1 Introduction

Employment, or rather the lack of it, that is to say, unemployment, is without doubt one of the most negative economic phenomena due to its potential consequences on the cohesion and stability of societies. Simply defined as “persons above a specified age not being in paid employment or self-employment but currently available for work during the reference period” (OCDE 1982), unemployment often reflects a mismatch between the number of job seekers and available job vacancies. Various reasons can create this mismatch. When there is not enough demand in the economy to provide jobs for everyone who wants to work (e.g., number of job seekers is larger than the number of openings), this is called cyclical (or Keynesian) unemployment. On the other hand, structural unemployment occurs when there is a mismatch between the skills of the unemployed workers and the skills needed for the available jobs.

Cyclical unemployment is largely affected by the status of the economy, which can significantly diminish job vacancies. Due to numerous recessions the world economy

✉ Nesrine Mezhoudi
nimezhoudi@iau.edu.sa

Rawan Alghamdi
rhalghamdi@iau.edu.sa

Rim Aljunaid
raaljunaid@iau.edu.sa

Gomathi Krichna
gKrishna@iau.edu.sa

Dilek Düşteğör
ddustegor@iau.edu.sa

¹ Computer Information Systems Department, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam 31441, Saudi Arabia

² Computer Science Department, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam 31441, Saudi Arabia

has experienced, the twentieth century is reported as a specific period for the emergence of global unemployment (Benjamin and Kochin 1982). While unemployment was very low in the beginning of the twentieth century, fluctuations were observed through two recessions in 1915 and 1921 (ILOSTAT database 2021). Following the great depression in the 1930s, numbers increased tremendously to reach the highest unemployment rates recorded after World War I (Benanav 2014). Following a time of recovery and steady state, the oil crisis in 1973 led to another recession between 1973 and 1975, followed by a period of almost minimal growth and again rising unemployment. The 1980–1982 recession, considered to be the most severe since world war II, marked the period where affected countries experienced high unemployment rates. Another severe depression took place in 1989, causing a recession in 1990–1991 whose effects lasted till 1994. Lately, in 2005, a gradual deterioration of the world economy and consequent slowed growth caused another wave of unemployment. Nowadays, it is expected that the COVID-19 crisis also affects employability. A study conducted by the Institute of Student Employers reveals that 27% of them would cut hiring this year (ISE 2020). Another 28% reported that they are currently uncertain of how the pandemic will affect their employment (Hooley 2020).

Major periods of structural unemployment also have been experienced in the same period of time, going through four industrial revolutions. The First Industrial Revolution (from 1760 to 1840) enabled transitioning to new manufacturing processes, moving from manual labor to machines thanks to the use of steam engine. The Second Industrial Revolution, also called technological revolution, was a period of rapid industrial development throughout Europe and USA, enabling mass production through assembly chain working. The Third Industrial Revolution, initiated by the progress in digital technology in the 1950s, caused the automation of the production with the introduction of industrial robots in factories. Throughout these industrial revolutions, specific expertise became obsolete almost overnight, thus creating hordes of useless workers and increased unemployment. Companies attempted to tackle this problem by providing training programs to qualify former employers and fresh graduates for the labor market. This was also followed by notable efforts and attempts from universities to bridge the gap between the outcomes of university education and the labor market by making fundamental changes to university systems, the programs offered, and the cooperative training mechanism (Layard et al. 1994). Finally, the Fourth Revolution, commonly called Industry 4.0, is the ongoing transformation of traditional practices thanks to latest progress in artificial intelligence (AI), using internet of things (IoT) infrastructure to provide self-monitoring, as well as

smart machines that can analyze and diagnose issues without the need for human intervention.

Amid the Industry 4.0 revolution, educational institutions and especially universities need to identify (or even anticipate) the new requirements of the labor market as impacted by ongoing transformations. Unfortunately, due to the simultaneous COVID 19 crises, companies are beginning to reduce the cost as much as possible by cutting out training applicants and searching for highly qualified graduates who can make an immediate contribution to the workplace once they are employed. This puts even more pressure on the educational systems whose graduates must cope with fierce competition. Developing tools and techniques to improve employability is thus of outmost necessity. For example, being able to identify the significant factors affecting employability, or the requirements of the evolving job market can tremendously help all stakeholders. Knowing their weaknesses and strengths, students might better plan their career. Instructors can focus on more appropriate skill sets to meet the requirements of rapidly evolving labor markets. Program managers can anticipate and improve their curriculum to build new competencies, both for educating, training and reskilling current and future workers. All these combined efforts certainly can contribute to overcome structural and cyclical unemployment, by increasing employability. Nonetheless to say that these tools must embed ethical and legal principles in their design, training, and deployment to ensure social benefits while still profiting from the promising potential of artificial intelligence (Ntoutsis et al. 2020).

The literature on graduates' employability shows that scientists started addressing this problem in the early 1980s. The first study, (Gisi and Forbes 1982), examined the results of the National Assessment of Educational Progress, which provides data on student achievement in various subjects, in combination with economic trends and future projections to reveal the shortcomings of students nationwide. Graduates' skills in mathematics, reading, science, and writing were assessed through a set of exercises that were manually evaluated by experts. Results were compared with percentages provided by the labor market. This method was notably labor intensive (thus both subjective and prone to human error), while also time consuming and costly.

We developed Fig. 1 by plotting the number of publications on employability every 5 years from 1980 to 2020. One can see that several studies have been conducted on studying the employability of graduates in various disciplines after the initial publication (Lynn Grover and Forbes 1982), with a noticeable increase after 1990, where studies roughly doubled every 5 years.

In Fig. 2, we differentiated the studies utilizing AI techniques vs other techniques, and in Fig. 3 we concentrated on AI related studies only displayed yearly. Figures 2 and 3 illustrate that most of these studies approached the problem

Fig. 1 Distribution of publications on employability per 5 years

Fig. 2 Percentage of publications using AI vs other methods

Fig. 3 Distribution of employability publications using AI techniques (as of August 2020)

using statistical methods to prove the relationship between different variables or make inferences from the current data, until 2010 when AI based methods started to be utilized. In 2020, there is a marked decrease in the number of studies on employability. However, note that these numbers were gathered in August 2020, thus they partially cover 2020. Also, 2020 was a special year due to Corona pandemic, following lock-down and various consequences in scientific publication and output.

There is an increase of AI based techniques. Yet, these studies show a lot of variation, for instance, with respect to how employability is defined, the data used, the machine

learning (ML) algorithms adopted, or even the research questions posed. Observing the heterogeneity of research applying diverse ML algorithms to answer different research questions related to employability prediction, there is a requirement to map and report on existing research, investigate significant models, examine case studies, gather statistical data, and make an inventory of best practices. In an attempt to depict a clear picture of the art, this study aims to compile the state-of-the-art and formulate answers to the following research questions based on the literature:

- How can employability be defined?
- What exactly can be predicted?
- What data can be mined?
- What AI methods / ML algorithms are used and are more performant?

These answers will provide a comprehensive roadmap enabling the application of data mining for employability prediction.

2 Review method

We focus on exploring studies addressing employability prediction throughout data driven and AI methods, motivated by the need to investigate and predict emerging employability factors. This survey brings together all works focused on predicting employability using ML techniques from different academic library; Web of Science, Scopus, and Google Scholar databases, and matching the machine learning based employability prediction topic. A total of relevant 20 papers were selected, that are using machine learning techniques to support employability prediction concerns. Selected studies were analyzed and compared with regards to the steps of data mining process. Further, we identified scenarios of employability predictions studies using ML and we outlines their common requirements and main challenges. In the remaining of this paper, first a summary of selected papers is provided in chronological order (Sect. 3). Then, a background section starts the analysis, where commonly used definitions for employability are summarized, along with the addressed problems in the literature (Sect. 4). In Sect. 5, we compiled the various machine learning methods used, with details on their datasets. Section 6 reports limitations and challenges associated with predicting employability.

3 Literature review

The earliest study aimed to enable higher education institutions to prepare their graduates to enter the job market with appropriate skills. For this, authors engrossed on factors

that influence the employability of the passed-out graduates, through survey data collected by the Ministry of Higher Education in Malaysia for the year of 2009 (Sapaat et al. 2011).

Later in 2013, authors develop a prediction model to predict the graduate's employability, using a three-scale categorization, namely employable, unemployable, and unpredictable, based on data resources from Maejo University in Thailand (Jantawan and Tsai 2013).

In a further attempt to identify students at risk of unemployment, (Mishra et al. 2016) develop a model by analyzing demographic, academic and emotional attributes of students from a Master of Computer Applications program in India. The same year, Piad, Keno et al. analyzed information technology student's demographic and cumulative grade point (CGPA) data from Philippine. This analysis helps to forecast the employability based on the present openings in the labor market (Piad et al. 2016).

Thakar et al. developed an employability prediction model using dataset of master's in computer applications students from various Colleges of a State University in Delhi, India, by considering academic attributes as main and psychometric attributes as secondary. The result showed students' employability prediction can be improved by including supplementary attributes such as personal, social, cognitive and other environmental variables (Thakar et al. 2017).

Rahman et al. broaden the scope by investigating different bachelor's degree (Science, Engineering, Psychology Educational Technology etc.), along with gender and CGPA of Malaysian University students. This study identifies pitfalls, and it supports university to produce graduates who meets the needs of job market (Rahman et al. 2017).

Bharambe et al. developed a model to map between company criteria and student skills. The assessment test data used to find their proficiency in various skill sets such as soft skills, problem-solving skills, technical skills, and weaknesses that they need to overcome for being employable in various companies. Student skillset data can also be helpful to map against the company's criteria (Bharambe et al. 2017).

García-Peñalvo et al. gathered data about employability parameters among the Spanish graduates after they leave the university. The model developed to identify the employment trends, and the analysis done to improve the future employment (García-Peñalvo et al. 2018).

Tapado et al. made significant contribution to identify supportive job training programs for the graduates to fit them as marketable and creates awareness about set of highly needed IT skills. And this study enables educational experts to make change in curriculum according to real world expectation (Tapado et al. 2018).

Othman et al. conducted a research on, the graduates of University Kebangsaan Malaysia (UKM). The seven

factors for better employment are researched namely their age, extra-curricular activities, field of study, marital status, communication skill, department and industrial training experience (Othman et al. 2018).

Wijayapala et al. conducted effective research on determining employability of university graduates in Sri Lanka. The predictive framework deals with four different domains namely employment status, job associated with degree, predicting the salary and field associated with job (Wijayapala et al. 2018).

Dubey and Mani developed a model for High school students of Loudoun County, VA, USA. This model predicts the employability of high school students with local businesses for part-time jobs (Dubey and Mani 2019). Alghamlas and Alabduljabbar conducted survey among bachelor's degree IT Students from Saudi Arabia universities and three months statistics poised from two online job gateways (LinkedIn and Bayt.com). This study indorses maximum desirable IT skills in the Saudi Arabia's employment market (Alghamlas and Alabduljabbar 2019).

Kumar developed a model using the dataset of engineering students from North India to predict the placement chances of graduates. This study helps the teachers and placement training centers to revise their curriculum and training methodologies for forthcoming years (Kumar 2019).

Denila et al. directed study among students from Davao del Norte State College, Philippines, and identify the correlation between students' academic grades in their core subjects and their first employment. Examining immediate employed graduates helps the university as well as students to progress in educating and producing valued graduates in coming years (Denila 2020).

Casuat and Festijo conducted research on Engineering Students from Technological Institute of the Philippines, the data analyzed are mock job Interview evaluation and survey conducted among students who were recently landed in job to identify most wanted skills (Casuat and Festijo 2019). The same dataset was later used to construct an employment prediction model, also investigating areas the student needs to progress to become more employable (Casuat et al. 2020).

Lately, Bhagavan et al. built a model using Hybrid Linear Vector Quantization (HLVQ) algorithm to analyze students' performance and used to predict chance of getting placed in desired organizations. This study aims to help the faculty progress in their research skills and to motivate the students to getting involved in research (Bhagavan et al. 2020).

Finally, Inger et al. and Patel et al. used the current job-ads in their prediction model. Patel et al. developed a deep learning model which also used as input the responses that students provided to a detailed questionnaire and quiz, assessed from the answers students' skill-levels that they finally mapped to job domain students' would be successful (Patel et al. 2020). Inger et al. on the other hand used job-ads

only and artificial intelligence methods to identify needs of the market (Mewburn et al. 2020).

In Table 1 summarizes the reviewed studies per level and domain of education. A more detailed analysis of these studies is provided in following sections.

4 Employability: background and context

The literature reveals several approaches to the problem of employability and its prediction, which differ per their definition of employability as well as the problem which they addressed (i.e., what is predicted). In the remaining of this section, the literature is categorized according to these perspectives. Definitions of employability are reported, as well as a summary of addressed problems.

4.1 Definition of employability

Despite a plethora of studies, the concept of employability is still debated by many scholars, and consensus has not been reached yet. Also, empirical research is insufficient to build a strong theoretical basis. For instance, when Harvey defines the employability as “the ability of the graduate to get a satisfying job” (Harvey 2001), his focus is on the satisfaction that the graduates get from his/her job. Similarly, for Hillage and Pollard “employability is the capability to move self-sufficiently within the labor market to realize

potential through sustainable employment” (Hillage and Pollard 1998), thus realizing one’s potential is important. On the same note, Lankard mentions that “employability skills as including personal image, interpersonal skills, and good habits and attitudes” (Lankard 1990), putting forward soft skills and especially relational skills. From a different perspective, when Harvey and Locke said, “it is the propensity of the graduate to exhibit attributes that employers anticipate will be necessary for the future effective functioning of their organization” (Harvey et al. 2002), they focus on the match between what graduates can bring and the demands. The European Association of Conservatories exhibits similar point of view in their definition as “the relevance of knowledge, skills and competences acquired through training to what the labor market/profession requires” (Association europeenne des conservatoires (AEC) 2004). The variation in views seems to mainly revolve around the focal point (e.g., the individual’s personality, his willingness to work, his skills).

Nevertheless, we can consider that there are two sides to employability, involving one’s skills set (both hard-skills and soft-skills) and the demand (defined by the market), which need to be matched. Relying on a causal dependency between these two sides, predictive methods are recently more and more used to estimate potential of a person with certain skills to meet the market demands. Machine learning based studies too, differ in their approach to employment, as they try to predict various dimension of the issue.

Table 1 Employability’ prediction studies: degree, field and country

Study	Level of education	Field	Country
Dubey and Mani (2019)	High School	Not specified	United States
Piad et al. (2016)	Bachelor	Information technology	Philippines
Rahman et al. (2017)	Bachelor	Science, engineering, psychology educational technology	Malaysia
Othman et al. (2018)	Bachelor	Not specified	Malaysia
Alghamlas and Alabduljabbar (2019)	Bachelor	Information technology	Saudi Arabia
Kumar (2019)	Bachelor	Engineering	India
Denila (2020)	Bachelor	Information technology	Philippines
Mishra et al. (2016)	Master	Computer applications	India
García-Peñalvo et al. (2018)	Degree and Master	Not specified	Spain
Thakar et al. (2017)	Bachelor and Master	Engineering, technology computer applications	India
Sapaat et al. (2011)	PhD	Not specified	Malaysia
Mewburn et al. (2020)	PhD	All	Australia
Jantawan and Tsai (2013)	Bachelor, Master, PhD	Computer science, information technology	Thailand
Bharambe et al. ((2017)	Not specified	Not specified	Not specified
Tapado et al. (2018)	Not specified	Information technology	Philippines
Wijayapala et al. (2018)	Not specified	Not specified	Sri Lanka
Casuat and Festijo (2019)	Not specified	Not specified	Philippines
Casuat et al. (2020)	Not specified	Not specified	Philippines
Bhagavan et al. (2020)	Not specified	Not specified	Not specified
Patel et al. (2020)	Not specified	Not specified	Not specified

4.2 Addressed problems

While studies on employability all aim to help reducing the problem of unemployment, we observe some differences on their specific objectives. We observe that ML based studies can be roughly categorized as follows:

- Predicting students' employability
- Skills mapping
- Adjusting curriculum
- Foreseeing long term market demand.

As one can see in Fig. 4, most studies aimed at identifying students' weaknesses, closely followed by predicting students' employability. All four categories are explained in detail in the following.

4.2.1 Predicting students' employability

An employability prediction model can be described as a function that takes as inputs some attributes related to the student and generates an output reflecting how "employable" this particular student is. Such models, if tailored for a particular job opportunity, could be used by HR specialists to pre-screen applicants and automatically eliminate mismatches. More generic models, tailored for a major, can help student advisors identifying students at risk and sign them up for reorientation or consulting services. Diverse application scenarios were implemented. Mishra et al. (2016), Tapado et al. (2018), and Bhagavan et al. (2020) motivated their research by the growing number of unemployed graduates. As a potential remedy, Bhagavan et al. (2020) aimed to predict students' employment chances from an educator's point of view, that is to say, they analyzed students' academic performance to assess their chance to get placed in reputed organizations. By the same token, Tapado et al. (2018) developed models to predict the score of each student

against employability based on their skills level. These models reveal strength and weakness of individual students before they appear for actual interviews. Similarly, Mishra et al. (2016) investigated the best suited algorithm to predict the employability of Master of Computer Applications (MCA) students.

4.2.2 Associating (skills mapping)

The studies under this category aim to identify if there is a gap in students' skills against companies' criteria. Skills identifications is an eye opener for students and academics about employability. Employability's associating studies enable a systematic assessment of job skills requirements, certification of student's skills, and mapping of graduate to certain positions and company. For instance, Thakar et al. (2017), Bharambe et al. (2017) focused their studies on graduate skills with an aim to support their chance for employability. The findings reported in Bharambe et al. (2017) support graduating students in bridging the gap between their profile and their desired company. Within the same perspective, Alghamlas and Alabduljabbar (2019) proposed a web-based application that assesses the suitability of IT students' skills for being recruited in the Saudi labor market. Further application examples are given in (Bhatia et al. 2018), where the proposed system mine large volume of students' data collected from their various usage of applications in the cloud. Students' academic skills are inferred and used to predict their employability at an early stage. Similarly, a web-based system is developed in Patel et al. (2020) where students answers to the questionnaire and quiz are used to assess their skill-sets and mapped to job domains that needs these skills.

4.2.3 Adjusting curriculum

Adjusting curriculum is a necessity, especially given the fast pace changes happening nowadays in the world economy. Thus, some of the ML based employability research are aimed to meet the requirements of the shifting landscape of higher education. This shift is emerging from the growing acknowledgement of graduate's unemployment crisis and the flaws in higher education developed programs. Mostly, the objective of these studies is to endorse the graduates' chances in employment by allowing them an education and skill development opportunities meeting job market requirements. Conducted research focused on adjusting curriculum, programs, and training definitions through promoting/demoting employability skills and considering identified requirements. For instance, Denila (2020)'s study argues for applying data mining techniques for assessing employment to support the university decisions in improving various education related aspects. Further, Rahman et al. (2017)

Fig. 4 Addressed problem and frequency of related articles

proposed a predictive model that will assist the university decision maker in creating better long-term plans for making future graduates more skilled, competent, and meet the industry requirements. Mewburn et al. (2020) showcased how to use natural language processing (NLP) methods to perform big data analysis on the text content of non-academic job advertisements and identified market needs for PhD holders in Australia. This way, they aimed to update curriculum to fit contemporary needs. Another important area of interest in application domains are, plan and execute skill set enhancement training programs according to students' level and efficient utilization of available resources. These techniques would help in overcoming the unemployment problem, while providing win-win strategy for students and academic institutions to get good placement ratio (Denila 2020).

4.2.4 Foreseeing long term market demand

Foreseeing long-term market demand can be defined as predicting future employment opportunities in the labor market by predicting factors that affect the demand of skills or competencies, as imposed by the company. These type of prediction models take information on factors affecting demand, then try to identify the most suitable relation among both independent and dependent variables (Green and Armstrong 2017). There have been some attempts in the literature, though not mature yet. Most of related studies have a rather short prediction window, which needs to be further expanded to talk about long term forecasting. For instance, Rahman et al. (2017) proposed a predictive model built using different algorithms for university's administration helping them in their long-term plans to produce graduates with advanced skills meeting the needs of the industry. However, their prediction window size is the shortest possible as models are used for fresh graduates. In Alghamlas and Alabduljabbar (2019), a web-based application is proposed designed to predict the suitability of IT students' skills for employment in the Saudi labor market and to provide a set of general recommendations regarding skills and competencies required to anticipate and prevent unemployment crisis. Their prediction window size is less than a year. Within the same perspective, Wijayapala et al. (2018) proposed an architectural framework which consists of four modules: employment status prediction, job salary prediction, job field prediction and job relevance prediction identifying considering the important factors relevant to each of four modules. The proposed framework aims to support students to predict their employment chances allowing them to enhance their skills at early stage to meet job market requirements in Sri Lanka. In this study, the used data belong to students that graduated two years ago, thus the prediction window size is two years. Furthermore, in 2018, a study in the information

technology domain proposed to find the important employability skills both from employers and student perspectives. The contrasting perception of skills helps to span the skill gap by finding out the skills that individuals possess and the skills that employer looks for at the time of recruitment which helps in predicting the future employment opportunities (Misra and Khurana 2018). Their prediction window size is one semester as predictive models are meant to predict senior students' employability.

5 Employability: frameworks and approaches

As mentioned in the introduction, unemployment is often the consequence of either sparsity of vacancies, or the gap existing between potential employees' developed skills and those that employers are seeking. The previous section reported how more and more recent studies utilize machine learning techniques to address unemployment issues. Success of these predictive models capitalize on diverse factors such as the data completeness, data size, machine learning techniques and, training process. Endorsing all those listed factors will result in a predictive model with enhanced accuracy. In this section, we follow the generic, well-known, and widely adopted data mining framework (Shearer 2000) to discuss the reviewed literature with respect to the technical factors supporting an accurate prediction, namely data acquisition, data preparation, model building, and model selection (see Fig. 5).

5.1 Data acquisition

An important factor for accuracy of predictive models is the data included in the process. It is important to deploy resourceful data for employability prediction, including various student information that can reveal various aspects of the relationship between students' attributes and employability skills. There are no studies demonstrating a clear linear and causal link between skills and attributes. Usually, researchers define a set of features to be the input of their predictive models, depending on what is available or easily accessible. Previous works have selected diverse dataset to assess employability from several sources: mainly universities, various related organisms (e.g., placement offices, career center office, ministry), and online job portals. But clearly, most of researchers considered students' academic records (Sapaat et al. 2011; Hogan et al. 2013; Thakar et al. 2017; Rahman et al. 2017; Bharambe et al. 2017; Tapado et al. 2018; Denila 2020), including data collected from placement reports which were considered valuable for the employability prediction (Piad et al. 2016; Othman et al. 2018; Kumar 2019; Bhagavan et al.

Fig. 5 Data mining process

2020). Some studies considered further details about the personal profile such as demographic profile, social academic integration, psychometric properties and emotional skills (Hogan et al. 2013; Mishra et al. 2016). Finally, very few studies use questionnaire data and quiz responses (Patel et al. 2020) as this is a time consuming process and the data is not readily available and large amount of data is necessary for training ML models. Besides students' information, also few studies utilize job-ads to identify the needed skills per job-domains (Patel et al. 2020; Mewburn et al. 2020).

Whether pre-university or university, academic records can easily be retrieved from the university Student Information System (SIS), nowadays used in most education institutions. SIS can also provide student demographics (e.g., age, gender, ethnicity), but socio-economic status might not be available explicitly. In that case, this could either be deduced from existing data, or it might be directly acquired from students through surveys (Wijayapala et al. 2018). However, psychological/soft skills data would probably need to be gathered directly from students. Accordingly, surveys were the second option for data acquisition, it allowed the collection of data in a structured way (Mishra et al. 2016; Bharambe et al. 2017; Tapado et al. 2018; Dubey and Mani 2019). Further, hybrid data acquisition method was considered where questionnaires' data and academic records were merged to build the dataset (Alghamlas and Alabduljabbar 2019; Casuat et al. 2020).

In Table 2 showcases for the ML based studies on employability prediction the features included, that we categorized as (1) hard skills (including academic performance), (2) soft skills, (3) demographics, (4) employment history, and (5) applied jobs. Hard skills were included in all the studies, followed by soft skills that were used in approximately 70% of studies, closely followed by demographics and employment history. Features related to applied jobs were used the least.

Among demographics/socio-economic features, gender was followed by parents' education, residence, parents' income, parents' job, and transportation. Having a student loan, ethnicity and age were only used in one study each (see Fig. 6). Overall, we observe that, except for gender, the most used features are related to the economic background of the students and his/her family.

Runner up after gender, the second most frequent features found in ML based studies are academic performance and IT skills (Fig. 7). Academic performance was measured using various indicators (e.g., grade point average, cumulative grade point average, or performance in a certain course), while IT skills was measured through programming languages, taken IT courses, or basic IT literacy depending on the background of the students involved in the study. Performance and IT skills are followed by problem solving/critical thinking skills, which is commonly cited among the twenty-first century skills. These skills refer to a broad set that are believed to be critically important to success in today's rapid changing world (Hogan et al. 2013; Wijayapala et al. 2018). They are not specific to a discipline (like IT, or computer science) and are reflecting the specific demands that will be placed upon graduates in a complex, competitive, knowledge-based, information-age, technology-driven economy and society. Other hard skills features used in the literature are degree, institution, number of courses taken, studentship (e.g., number of study hours per day, GAP year, way of studying the subject, attendance), and graduate studies (Fig. 7).

Communication skills also were used as often as academic performance and IT skills to predict employability of students. Other soft skills used were psychometric aptitude, general soft skills, leadership, self-esteem, appearance, socializing, alertness, personality, assertion, empathy, decision making, time management, stress management, and teamwork. Nevertheless, except communication skills, other soft skills were not included very often, probably due to the difficulty to measure them (Fig. 8).

Table 2 Employability predicting features supported by different approaches

	Hard Skills	Soft skills	Demographics	Employment history	Applied job
Bharambe et al. (2017)	•	•			
Thakar et al. (2017)	•	•	•		
Alghamlas and Alabduljabbar (2019)	•	•			
Rahman et al. (2017)	•		•		
Tapado et al. (2018)	•			•	•
Denila (2020)	•				
(Mishra et al. (2016)	•	•	•	•	
Wijayapala et al. (2018)	•	•	•	•	•
Dubey and Mani (2019)	•	•	•	•	
García-Peñalvo et al. (2018)	•		•		
Casuat et al. (2020)	•	•		•	
Piad et al. (2016)	•		•		
Sapaat et al. (2011)	•	•	•	•	
Jantawan and Tsai (2013)	•	•	•	•	•
Othman et al. (2018)	•	•	•	•	
Kumar (2019)	•	•	•		
Casuat and Festijo (2019)	•	•			
Bhagavan et al. (2020)	•				
Patel et al. (2020)	•				•
Mewburn et al. (2020)					•

Fig. 6 Details and frequency of demographics included in related papers

Employment type was used in one third of studies, to indicate if past employment experience were permanent, contractual, freelancer, consultant or part-time. Other features related to employment history were employment industry, reason of unemployment, job category, position, salary, time to find work, job relevance to obtained education, problems in job, date of employment, and job location. Similar to most skills features, these were rarely included in the prediction models (Fig. 9).

Fig. 7 Details and frequency of hard skill features in related papers

Finally, very few studies included features related to sought employment. Job relevance (alignment with obtained degree) was used in two studies, expected salary, expected sector, location and position were used only in one study. Also, one study that collected data from questionnaire, asked what matters in accepting a job, with following answers, prestige of the employer, tasks to be done, position, and family (Fig. 10).

Fig. 8 Details and frequency of soft skills included in related papers

Fig. 9 Details and frequency of employment history included in related papers

Fig. 10 Details and frequency of applied job included in related papers

5.2 Data preparation

Once collected, the raw data need to undergo a preparation phase and pre-processing to be ready for analysis and modeling. The dataset, that is mostly obtained from merging various tables, might contain missing data, inconsistent data,

incorrect data, miscoded data, and duplicate data. Thus, data preparation and pre-processing improve the quality of the dataset and consequently reduce potential erroneous conclusions.

Data preparation and pre-processing usually include several steps, namely (1) selection, (2) cleaning, (3) derivation of new variables, (4) data transformation, (5) handling imbalanced data sets, and (6) feature selection. Details on all these stages can be found in Alyahyan and Düşteğör (2020) and there in references. In the following, we mention the steps that were reported in the selected literature.

The dimension of the data gathered can be significant, especially if complete transcripts information is included. Large number of features, especially with a lower number of samples, creates what is called the curse of dimensionality, negatively impacting both the computation time and the quality of prediction (Tatar and Düşteğör 2020). Therefore, data selection is necessary to identify important attributes (García-Peñalvo et al. 2018). For example, (Mishra et al. 2016; Thakar et al. 2017; Rahman et al. 2017; García-Peñalvo et al. 2018; Tapado et al. 2018; Saouabi and Abdellah 2019; Casuat et al. 2020; Denila 2020) applied data selection to remove non-pertinent parameters such as student name, and institute name.

Since cleaning data avoids inconsistency, noise, and enables to handle missing values, it is important to know how to perform this step without compromising the quality of the prediction. Many of the studies included in the survey referred to this step, as missing data is a common fact when data is collected through questionnaire (Sapaat et al. 2011; Jantawan and Tsai 2013; Mishra et al. 2016; Thakar et al. 2017; Rahman et al. 2017; García-Peñalvo et al. 2018; Tapado et al. 2018; Saouabi and Abdellah 2019; Casuat et al. 2020; Denila 2020). Academic data also can show these problems due to irregular and transfer students. (Casuat et al. 2020) managed missing data by replacing them with corresponding median values in order to preserve statistical properties of the dataset.

New variables sometimes can be derived from existing variables, which can improve the predictive model (Nisbet et al. 2009). For example, Piad et al. (2016) transform students' address into a binary location attribute, indicating whether they are leaving in rural or urban areas.

Data transformation is a necessary process to eliminate dissimilarities in the dataset, it includes normalization of numeric attributes (necessary when the data includes varying scales), discretization (necessary when using data mining techniques that allow only for categorical variables), and conversion to numeric variables (as most ML algorithms offer better results using a numeric variable). García-Peñalvo et al. (2018) considered the one-hot encoding algorithm to normalize the dataset attributes into categorical values. In Piad et al. (2016), attributes were discretized into intervals

with categorical or nominal attributes to enable their classification. Similarly, Sapaat et al. (2011, Bhagavan et al. (2020) applied discretization for six attributes with continuous value (gpa, age ...).

For some particular problem, it is common that the dataset is imbalanced, in our case meaning that the number of unemployed graduates is significantly more than employed graduates. The concept of imbalanced data is frequently found in real-world data and have negative impacts on the quality of the predictive model and its accuracy. Several solutions were identified to mitigate imbalanced class problem, like under sampling, or over sampling. Despite the importance of balancing data, reviewed papers do not provide any details in this regard except in Wijayapala et al. (2018).

When the data set is prepared and ready for modeling, then the important variables can be chosen and submitted to the modeling algorithm. This step, called feature selection, aims to choose a subset of attributes from the input data with the capability of giving an efficient description for the input data while reducing effects from unrelated variables. There exist several methods, like filter, wrapping, and embedded methods, but all require trial and error which can be a tedious process. Authors in Wijayapala et al. (2018) applied information gain which promoted the job sector attribute for predicting employability. Othman et al. (2018) also used

information gain to select attributes that are the most important, and selected (Age, Faculty, Field, Family Income, Co-curriculum, Marital Status, Industrial Internship, English Language Skill, Employability Status).

Despite the importance of the preprocessing phase, some works do not provide enough details about applied techniques (Piad et al. 2016; Bharambe et al. 2017; Othman et al. 2018; Alghamlas and Alabduljabbar 2019; Kumar 2019; Patel et al. 2020; Mewburn et al. 2020). A detailed description of preprocessing practices in the related literature is presented in Table 3.

5.3 Model building

Analyzing the related works presented in this study, we notice two approaches utilized for employability prediction approaches, namely data mining and machine learning.

While data mining is a subset of business analytics, it refers to exploring an existing large dataset to discover previously unknown patterns, relationships and anomalies that are present in the data. Thus, it is used to find new insights on the data. Machine learning, on the other hand, is a subset of artificial intelligence (AI), where computers analyze large datasets to 'learn' patterns that will making predictions for new data. However, the frontier between both approaches is blurry, as machine learning may use some data mining

Table 3 Preprocessing practices supported by different approaches

	Selection	Cleaning	Derivation of new variables	Data transformation	Imbalanced data sets	Feature selection
Bharambe et al. (2017)						•
Thakar et al. (2017)		•		•	•	•
Alghamlas and Alabduljabbar (2019)						•
Rahman et al. (2017)		•				•
Mishra et al. (2016)		•				•
Tapado et al. (2018)		•				
Casuat and Festijo (2019)	•			•		•
Denila (2020)	•	•				
Kumar (2019)		•				•
Wijayapala et al. (2018)					•	
Dubey and Mani (2019)						•
García-Peñalvo et al. (2018)	•	•		•		
Casuat et al. (2020)	•			•		•
Piad et al. (2016)			•			•
Sapaat et al. (2011)		•		•		•
Jantawan and Tsai (2013)		•		•		•
Othman et al. (2018)		•				
Bhagavan et al. (2020)		•		•		
Patel et al. (2020)			•			
Mewburn et al. (2020)			•			

techniques to build models and find patterns, so that it can make better predictions. Data mining also can sometimes use machine learning techniques to produce more accurate analysis. Thus, in this study we look at both approaches interchangeably.

Machine learning uses two types of techniques: unsupervised learning, which finds hidden patterns in input data without labeled responses, and supervised learning, which trains a model on known input and output data so that it can predict future output. Unsupervised learning has the potential to reveal interesting unknown relationship or patterns. However, in this study, since we are interested in employability prediction, that is to say, we have a well-defined target variable, supervised learning was considered. Commonly identified studies in the literature deployed supervised learning techniques. Supervised learning uses classification and regression techniques to develop predictive models. While regression techniques predict continuous responses, classification techniques predict discrete responses. Since regression would need the ability to assign a continuous score of employability, all studies define the employability prediction problem as a classification problem. Classification models classify input data into categories. Most studies define two classes, namely *employable*, and *unemployable*; some studies also define a third class to represent students that are in-between, as *unknown*.

Despite the diversity of classification methods, the studied literature reveals that mostly decision tree, random forest, support vector machine, naïve Bayes and nearest neighbor were applied on the data to predict the employability of students and/or to determine the factors impacting the employability of students (Hogan et al. 2013; Mishra et al. 2016; Bharambe et al. 2017; Wijayapala et al. 2018; Tapado et al. 2018; Bhagavan et al. 2020). With the recent growing interest in deep learning, Patel et al. (2020) used multi-layer neural network to assess students' answer to a comprehensive questionnaire and quiz, then mapped the answers to most appropriate job-domain. A further detailed prediction process was presented by Saouabi and Abdellah (2019), which consists of an intelligent system using data mining techniques on employability data in a Big Data environment. The system represents the prediction process through seven stages and defines the appropriate tool and techniques to execute each phase (Saouabi and Abdellah 2019). Piad et al. (2016) considered the process of knowledge Discovery in Databases (KDD) to predict employability. Three phases of the data mining techniques were selected, including data pre-processing, classification tasks, interpretation, and evaluation.

The analysis of previous works also shows a common practice of performing a comparative evaluation of different ML algorithms to evaluate their performance. However, their conclusions and findings are not congruent. For instance

(Piad et al. 2016) compared five classification algorithms applied for an IT employability dataset. The results show that the highest accuracy was obtained by logistic regression compared to Chaid, Naïve-Bayes, J48, and SimpleCart (Piad et al. 2016). However, results in Sapaat et al. (2011) showed that J48, a variant of the decision-tree algorithm, produced the highest accuracy compared to Bayes algorithms and several tree-based algorithms (Sapaat et al. 2011). Kumar (2019) build their classifier using Decision Tree, Support Vector Machine, Gaussian Naïve Bayes, and K-Nearest Neighbor. Their results show that the Decision Tree outperformed all other methods. All these conflicting results can be explained with the fact that datasets have varying characteristics, such as dimensionality, low volume of data. Thus, predictive models are usually tailored for a certain data set.

Some studies also opted to integrate more than one ML technique at different stages of the prediction process. For instance (Thakar et al. 2017) proposed a unified model to automate the selection of relevant attributes from the set of the dataset. Along with a classification model integrating four best classifiers with vote ensemble method to predict student's employability (Thakar et al. 2017).

5.4 Model selection

The model selection phase is the common challenge of applied machine learning. It is about identifying and selecting the appropriate choice among a range of different models that could be used for the problem. While model accuracy is an important factor, other concerns like how long the model takes to train (complexity), or how easy it is to explain to stakeholders (interpretability) are also important.

Existing works assessed and compared the performance of tested models and proposed models using different measures such as accuracy, F1-score, precision, recall, ROC area, root mean square error, root relative square error, and error rate. In their work, Casuat and Festijo (2019 and Casuat et al. (2020), authors evaluated the performance of models and compared with regards to accuracy, recall and precision and F1 score measures. The performance of the model in Thakar et al. (2017) is measured by classification accuracy, kappa, and F1 Score.

Further, to minimize variance, and make sure that recorded performances are genuine, the n-fold cross-validation techniques were applied in most studies (Bhatia et al. 2018; Mishra et al. 2016; Thakar et al. 2017; Bharambe et al. 2017; Wijayapala et al. 2018; Tapado et al. 2018; Casuat and Festijo 2019; Kumar 2019). This consists of keeping an independent test dataset, that the model withholds from during training and model selection, to avoid the leaking of test data in the training stage. Tapado et al. (2018) applied a cross Validation with tenfolds test option on IT Job Category and selected J48 Decision Tree Algorithm. We noted that

the same algorithm was selected in diverse situations for different goals based on diverse measures as detailed in the table Y (Sapaat et al. 2011; Jantawan and Tsai 2013; Mishra et al. 2016; Piad et al. 2016; Othman et al. 2018). Aiming at enhancing accuracy, Dubey et al. (2019) used bootstrapping in order to increase the size of the dataset to 195 records.

Interpretability is a significant capacity of ML algorithm, ideally sought in predictive models. An interpretable model, besides its predictive nature, also pinpoints which attributes are the most important towards employability. Supporting interpretability reflects the importance of allowing stakeholders to comprehend why certain decisions or predictions have been made and increase the trust. Among the ML algorithms cited in this survey, only the Decision-tree J48 is recognized to be interpretable along with Decision Tree-C4.5, Logistic Regression and Bayesian methods. The deployment of these methods allowed to identify the important factors and determinants of employability (Kumar 2019; Mishra et al. 2016; Tapado et al. 2018; Denila 2020; Dubey et al. 2019; Piad et al. 2016; Causat et al. 2019; Sapaat et al. 2011; Othman et al. 2018; Jantawan and Tsai 2013). For instance, Othman et al. (2018) identified the most influential attribute in classifying employment and unemployment classes; in their study, employment is determined by the age attribute, industrial internship, English skills and involvement in curriculum activity. Meanwhile, unemployment relates to the marital status and field of study. In Mishra et al. (2016), authors found that Empathy, Drive and Stress Management abilities are the major emotional parameters that affect employability. Finally, in Piad et al. (2016), authors identified three academic variables as significant predictors for employability namely, IT Core skills along with the gender. Table 4 summarizes for each study their performance criteria used to assess and compare various ML methods. Highlighted models are interpretable.

6 Employability prediction: challenges, and limits

Predicting employability using data mining techniques implies several challenges and limitations that need to be overcome. This section summarizes the challenges and limitations commonly encountered in the literature.

The major challenge for employability studies is collecting consistent and quality data. Often, ML algorithms perform better with larger datasets, as well as with more information about the student / future employee. For example, in Rahman et al. (2017), the authors mention the need for gathering more attributes such as grade subjects taken during the study period, the results of the oral test and work status to have a complete dataset. Unfortunately, collecting

extra information can be expensive (especially if it is not already available in the school records), and sometimes even legally challenging due to privacy reasons.

Further, we found that the most used feature in all studies was found to be gender. This is not a surprise as gender gap in employment is a worldwide fact, that many governments and organizations are now actively addressing as they seek gender parity. Surprisingly, only few studies considered psychometric attribute (Thakar et al. 2017) despite their role in employability, as was recently pointed out (Potgieter et al. 2020; Roslan et al. 2020). These recent studies showed that psychometric attributes contribute to the analysis of youth unemployment issues. On the other hand, no paper considered the economical context and features related with the country economy despite their significance for the job market and employment. We also observe that none of the studies collected CVs, and only two studies collected job-ads (Patel et al. 2020; Mewburn et al. 2020), although these are the two main inputs traditionally used by recruiting agents and job seekers.

One more challenge is about the completeness and accuracy of gathered data. This is especially the case when the data collected via web forms as the graduate can quit the web form in any moment (García-Peñalvo et al. 2018), or via e-learning activities (Bhatia et al. 2018). Other studies which faced the same issue are (Sood and Singh 2019) which had some missing values after the data is stored in a cloud repository. Casuat et al. (2020) solved the issue by filling the missing values with the median value. Piad et al. (2016) had some missing values and outliers as well. If not handled properly, missing value can compromise the quality of prediction, especially for Support Vector Machines, Neural Networks, Naive Bayes, and Logistic Regression (Alyahyan and Düşteğör 2020). Although the lack of general procedure that can be applied to all problems, there exist several methods to address this issue. As explained in the Sect. 4.2, this is commonly done as preprocessing of the data in the preparation phase.

As a consequence of the nature of the problem, it is likely to have an imbalanced dataset where the majority class is employable (Thakar et al. 2017; Casuat et al. 2020). Again, this lack of balance may negatively impact the performance of data mining algorithms, and re-sampling (under or over-sampling) is the solution of choice as explained in Sect. 4.2.

Another challenge, also associated with ML methods, is to decide which prediction model is good enough. Building a best prediction model includes several factors, like which features to select, which ML algorithm to use, or for a given algorithm, which hyper-parameters to select. While an exhaustive search would yield an undeniable best model, due to its size, searching the whole parameter space would often become computationally intractable. Heuristic search methods are often used instead. But none of the studies in the literature reported using such methods.

Table 4 Comparison of supported measures to assesses and compares the performance of prediction-models (where, ACC: accuracy; t: time to build model; Prec: precision; RMSE: root mean square error; RRSE: root relative square error; highlighted models are interpretable)

Selected model	Performance Measures									
	ACC	T	F1-score	Prec	Recall	RMSE	Error rate	ROC Area	RRSE	
Bharambe et al. (2017)	•	•					•			
Thakar et al. (2017)	•		•	•	•					
Rahman et al. (2017)	•					•	•			
Mishra et al. (2016)		•	•					•		
Tapado et al. (2018)	•					•		•		
Casuat et al. (2020)	•		•	•	•					
Denila (2020)	•		•	•	•					
Wijayapala et al. (2018)	•		•	•	•					
Dubey and Mani (2019)	•		•	•	•					
García-Peñalvo et al. (2018)			•	•	•					
Othman et al. (2018)	•	•	•			•		•		
Piad et al. (2016)	•						•			
Sapaat et al. (2011)	•					•	•			
Jantawan and Tsai (2013)	•					•	•		•	
Kumar (2019)	•			•				•		
Casuat and Festjo (2019)	•		•	•	•					
Alghamlas and Alabduljabbar (2019)										
Bhagavan et al. (2020)	•	•	•	•	•					
Patel et al. (2020)	•									

Scalability is also something that needs to be correctly understood, thus reported results need to be handled with care. Often, studies are conducted for a specific group of graduates, which makes their results hardly generalizable to a more diverse population. However, if future empirical studies indicate no difference in skill requirements between communities, researchers could try to generalise study results to a wider population.

Indeed, scalability is achieved when conducting a survey with a sample of large cohort that represents the "whole" we want to draw conclusions about. Conducting a survey this way will assure to have a generalizable study with scalable results. We observe however that the studied literature often focusses on one cohort. For example, Casuat et al. (2020) and Denila (2020) are restricted to the same university or college. On the other hand, Mishra et al. (2016), García-Peñalvo et al. (2018), Alghamlas and Alabduljabbar (2019), used a mixed cohort from various universities/colleges, thus their results can be generalized to a more diverse population. Figure 11 illustrates how scalable are the studies, whether the study has been done on a diverse colleges/institutes or only on one college/institute. Some studies did not specify the domain of their studies.

Closely related to scalability is the issue of reproducibility. Reproducibility is a challenge in applied data science and produces a lot of discussion (Molner Domenech and Guillén 2020). First conditions for a study to be reproducible is the availability of dataset as well as source code. If a study is not reproducible, results need to be treated carefully. None of the

studies included in this review publicly shared their datasets or their source code, except pseudocode for one study only (Bhagavan et al. 2020).

7 Discussion and conclusion

We identified four main objectives in the studies on data mining application for employability prediction, namely classifying students as employable or not, identifying students' lacking skills, adjusting curriculum, and foreseeing long term market demand. All studies aiming to assess students' employability reached promising accuracy demonstrating that using machine learning techniques for employability prediction is feasible. Summarizing here significant lacking skills or important determinants of employment: gender and score obtained in IT core courses were found to be important factors among bachelor students from an IT department in Philippines (Piad et al. 2016). Still in the field of IT, most needed skills in the IT job market in Saudi Arabia were identified as hard skills (especially development), soft skills (mainly, communication), and programming languages (specifically, JavaScript). The same study also pinpointed a lack in educational output regarding the hard skills among young graduates (Alghamlas and Alabduljabbar 2019). In a related field, namely computer application, working knowledge of Java and C#, along with stress management were identified as mainly sought by recruiters in India. Important determinants identified among employed students were academic hours put into study, and empathy.

Fig. 11 Studies scalability

Students living in cities (as opposed to countryside) were also more likely to get employable, especially if they got an education loan (Mishra et al. 2016). Age, along with scores in English, mathematics and logics were found to be significant attributes in predicting employability of Bachelor and Master graduates from School of Engineering, Technology and Computer Applications in India (Thakar et al. 2017). Among BS graduates from an engineering department in India, CGPA has been found to be the most important factor affecting employment. In the Malaysian study spanning several departments and colleges, age, industrial internship and students' department were found to contain the most information and affect the final class, i.e., employability status (Othman et al. 2018). In terms of curriculum adjustment, (Tapado et al. 2018) suggested several policies modifications to enhance the employability of the IT graduates, like enabling industry immersion and enhancing IT skills through new courses. They also suggested upgrading IT facilities and infrastructure, limiting class capacities, as well as ensuring the faculty body is UpToDate with latest technologies. The body of related studies shows significant variations both in their aims and findings, as we clearly demonstrated in this paper.

We listed all students' attributes used and classified them as hard skills, soft skills, demographics, employment history, and applied job. We reminded the main preprocessing steps, especially highlighting the probable issues in the context of prediction of employability. Table 3 clearly shows that existing studies are either missing most of these important preparation steps or failed to report them. We reported all machine learning algorithms that were used and compared to predict employability; reporting that deep learning methods are rarely used although once again identifying the weakness in searching for the optimal parameters, as none of the publications reported this step. Even cross validation, that is commonly practiced to minimize variance and to guarantee sound performance results, was found to be rarely reported. Talking of performance, various metrics have been found in the literature (e.g., accuracy, F1-score, precision, recall, ROC area, root mean square error, root relative square error, and error rate), however no explanation was provided on which measure is more appropriate to the problem of employment prediction.

While pinpointing their divergences and identifying weaknesses of the body of studies, we conducted a thorough and systematic analysis that we aligned with the main steps of data mining process, thus putting forward a common story line. For future studies, we recommend justifying various decisions with respect to the context of the specific objectives and research questions. Following the comprehensive roadmap presented in this study, undergoing all steps (especially the steps identified here as missing or weak) will yield enhanced, generalizable, and interpretable results.

Acknowledgements Authors would like to recognize and thank the precious support from the Deanship of Scientific Research at Imam Abdulrahman bin Faisal University.

Author contributions Conceptualization DD. Methodology DD. Formal analysis NM. Investigation NM, RaA, RA, GK. Resources NM, RaA, RA, GK. Writing—original draft preparation NM, RaA, RA, GK. Writing—reviewing—editing NM, RaA, RA, GK, DD. Visualization DD, NM, RaA. Supervision DD. Project administration DD.

Funding This study was supported by the Deanship of Scientific Research (DSR), Imam Abdulrahman Bin Faisal University (IAU) under the project number: 2019-387-CSIT.

Declarations

Conflict of interest The authors declare that they have no competing interests.

References

- Alghamlas M, Alabduljabbar R (2019) Predicting the suitability of IT students' skills for the recruitment in Saudi Labor Market. In: 2nd International Conference on Computer Applications and Information Security, ICCAIS 2019. Institute of Electrical and Electronics Engineers Inc
- Alyahyan E, Düstegör D (2020) Predicting academic success in higher education: literature review and best practices. *Int J Educ Technol High Educ* 17:3
- Association europeenne des conservatoires (AEC) (2004) Glossary of terms used in relation to the Bologna Declaration. <https://www.qualityresearchinternational.com/glossary/employability.htm>. Accessed 12 Sep 2020
- Benanav AS (2014) A global history of unemployment: surplus populations in the World Economy, 1949–2010. UCLA. ProQuest ID: Benanav_ucla_0031D_13186. Merritt ID: ark:/13030/m5253s28
- Benjamin DK, Kochin LA (1982) Unemployment and unemployment benefits in twentieth-century britain: a reply to our critics. *J Polit Econ* 90:410–436
- Bhagavan KS, Thangakumar J, Subramanian DV (2020) Predictive analysis of student academic performance and employability chances using HLVQ algorithm. *J Ambient Intell Humaniz Comput*. <https://doi.org/10.1007/s12652-019-01674-8>
- Bharambe Y, More N, Mulchandani M et al (2017) Assessing employability of students using data mining techniques. In: 2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017. Institute of Electrical and Electronics Engineers Inc, pp 2110–2114
- Bhatia M, Sood S, Sood SK (2018) Exploring temporal analytics in fog-cloud architecture for smart office healthcare internet of things view project an automatic prediction of epileptic seizures using cloud computing and wireless sensor networks view project exploring temporal analytics in fog-cloud architecture for smart office healthcare. <https://doi.org/10.1007/s11036-018-0991-5>
- Casuat CD, Festijo ED (2019) Predicting students' employability using machine learning approach. In: ICETAS 2019 - 2019 6th IEEE International Conference on Engineering, Technologies and Applied Sciences. Institute of Electrical and Electronics Engineers Inc
- Casuat CD, Festijo ED, Alon AS (2020) Predicting students' employability using support vector machine: a smote-optimized machine

- learning system. *Int J Emerg Trends Eng Res* 8:2101–2106. <https://doi.org/10.30534/ijeter/2020/102852020>
- Denila PG (2020) Analysis of IT graduates employment alignment using C4.5 and Naïve Bayes algorithm. *Int J Adv Trends ComputSciEng* 9:745–752
- Dubey A, Mani M (2019) Using machine learning to predict high school student employability—a case study. In: *Proceedings—2019 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2019*. Institute of Electrical and Electronics Engineers Inc, pp 604–605
- García-Peñalvo F, Cruz-Benito J, Martín-González M et al (2018) Proposing a machine learning approach to analyze and predict employment and its factors. *Int J Interact MultimedArtifIntell* 5:39. <https://doi.org/10.9781/ijimai.2018.02.002>
- Gisi L, Forbes R (1982) The information society: are high school graduates ready? Education Commission of the States, Denver, CO
- Green KC, Armstrong JS (2017) Demand forecasting: evidence-based methods. *SSRN Electron J*. <https://doi.org/10.2139/ssrn.3063308>
- Harvey L (2001) Defining and measuring employability. *Qual High Educ* 7:97–109. <https://doi.org/10.1080/13538320120059990>
- Harvey L, Locke W, Morey A (2002) Enhancing employability, recognising diversity Universities UK. London
- Hillage J, Pollard E (1998) Employability: developing a framework for policy analysis|VOCEDplus, the international tertiary education and research database. Suffolk
- Hogan R, Chamorro-Premuzic T, Kaiser RB (2013) Employability and career success: bridging the gap between theory and reality. *Ind Organ Psychol* 6:3–16. <https://doi.org/10.1111/iops.12001>
- Hooley T (2020) Redefining the employability agenda in the age of Covid-19 - HEPI. In: *High. Educ. Policy Inst.* <https://www.hepi.ac.uk/2020/04/11/redefining-the-employability-agenda-in-the-age-of-covid-19/>. Accessed 17 Sep 2020
- ISE I of SE (2020) COVID-19: Challenges for student recruitment - Institute of Student Employers | ISE. <https://ise.org.uk/page/blog-covid-19-challenges-for-student-recruitment>. Accessed 25 Sep 2020
- Jantawan B, Tsai C (2013) The application of data mining to build classification model for predicting graduate employment. *Int J Comput Sci Inf Secur* 11(10):1–7
- Kumar P (2019) Anticipating placement status of students using machine learning. *J Gujarat Res Soc* 21(6):738–747
- Lankard BA (1990) Employability—The Fifth Basic Skill. *ERIC Digest* No. 104
- Layard R, Nickell SJ, Jackman R (1994) *The unemployment crisis*. Oxford University Press, Oxford, UK. ISBN 019877395
- Mewburn I, Grant WJ, Suominen H, Kizimchuk S (2020) A Machine learning analysis of the non-academic employment opportunities for Ph.D. graduates in Australia. *High Educ Policy* 33:799–813. <https://doi.org/10.1057/s41307-018-0098-4>
- Mishra T, Kumar D, Jambheshwar G (2016) Students' performance and employability prediction through data mining: a survey. *Int J Appl Eng Res* 11:2275–2282. <https://doi.org/10.17485/ijst/2017/v10i24/110791>
- Misra RK, Khurana K (2018) Analysis of employability skill gap in information technology professionals. *Int J Hum Cap InfTechnol Prof* 9:53–69. <https://doi.org/10.4018/IJHCITP.2018070104>
- Molner Domenech A, Guillén A (2020) MI-experiment: a Python framework for reproducible data science. In: *Journal of Physics: Conference Series*. 12025
- Nisbet R, Elder J, Miner G (2009) Chapter 11—Classification. *Handbook of statistical analysis and data mining applications*. Academic Press, Boston, pp 235–258
- Ntoutsi E, Fafalios P, Gadiraju U et al (2020) Bias in data-driven artificial intelligence systems—an introductory survey. Wiley
- Interdiscip Rev Data Min KnowlDiscov. <https://doi.org/10.1002/WIDM.1356>
- OCDE (1982) International Labour Organization (ILO) resolutions concerning economically active population, employment, unemployment and underemployment. In: *Thirteenth International Conference of Labour Statisticians*
- Othman Z, Shan SW, Yusoff I, Kee CP (2018) Classification techniques for predicting graduate employability. *Int J Adv Sci Eng Inf Technol* 8:1712–1720. <https://doi.org/10.18517/ijaseit.8.4-2.6832>
- Patel A, Mascarenhas S, Thomas A, Varghese D (2020) Student performance analysis and prediction of employable domains using machine learning. *SSRN Electron J*. <https://doi.org/10.2139/ssrn.3682499>
- Piad KC, Dumlaio M, Ballera MA, Ambat SC (2016) Predicting IT employability using data mining techniques. In: *2016 3rd international conference on digital information processing, data mining, and wireless communications, DIPDMWC 2016*. Institute of Electrical and Electronics Engineers Inc, pp 26–30
- Potgieter IL (2020) Personality and psycho-social employability attributes as meta-capacities for sustained employability. In: *Psycho-social career meta-capacities*. Springer, Cham, pp 35–54
- Rahman NAA, Tan KL, Lim CK (2017) Predictive analysis and data mining among the employment of fresh graduate students in HEL. *AIP Conf Proc* 1891:020007
- Roslan S, Ping WS, Sulaiman T, Jalil HA, Siaw YL (2020) Psychometric properties of graduate employability instrument among Malaysian higher education institution students. *Univ J Educ Res* 8(1A):164–178
- Saouabi M, Abdellah E (2019) Proposition of an employability prediction system using data mining techniques in a big data environment. *Int J Math Comput Sci* 14(2):411–424
- Sapaat MA, Ai M, Ahmad J et al (2011) A data mining approach to construct graduates employability model in Malaysia. *Int J New ComputArchit Their Appl* 1:1086–1098
- Shearer C (2000) The CRISP-DM model: the new blueprint for data mining. *J Data Warehouse* 5(4):13–22
- Sood SK, Singh KD (2019) Optical fog-assisted smart learning framework to enhance students' employability in engineering education. *ComputApplEngEduc* 27:1030–1042. <https://doi.org/10.1002/cae.22120>
- Tapado BM, Acedo GG, Palaoag TD (2018) Evaluating information technology graduates employability using decision tree algorithm. In: *ACM international conference proceeding series*. Association for Computing Machinery, New York, USA, pp 88–93
- Tatar AE, Düşteğör D (2020) Prediction of academic performance at undergraduate graduation: course grades or grade point average? *Appl Sci* 10(14):4967
- Thakar P, Mehta A, Manisha (2017) A unified model of clustering and classification to improve students' employability prediction. *Int J IntellSystAppl* 9:10–18. <https://doi.org/10.5815/ijisa.2017.09.02>
- Unemployment, total (% of total labor force) (modeled ILO estimate). International Labour Organization, ILOSTAT database. Data retrieved on January 29, 2021. <https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS>
- Wijayapala MP, Premaratne HL, Jayamanne IT (2018) Employability and related context prediction framework for university graduates: a machine learning approach. *Int J Adv ICT EmergReg* 9:1. <https://doi.org/10.4038/ictcr.v9i2.7181>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.