Revista Industrial Data 25(2): 275-300 (2022)
DOI: https://doi.org/10.15381/idata.v25i2.21643.g19078

ISSN: 1560-9146 (Impreso) / ISSN: 1810-9993 (Electrónico)
Facultad de Ingeniería Industrial - UNMSM

Machine Learning Process to Determine the Social Demand for IT Professional Jobs

ZORAIDA MAMANI RODRIGUEZ [1]

SUBMITTED: 06/12/2021 ACCEPTED: 12/01/2022 PUBLISHED: 31/12/2022

ABSTRACT

Machine learning is a branch of artificial intelligence that uses scientific computing, mathematics and statistics through automated techniques to solve problems based on classification, regression and clustering. Social demand refers to the need for service and product of the professional training process, expressed by interest groups, aimed at contributing to national development, as established by the quality assurance policy of university higher education and national licensing and accreditation models. In this context, this paper conducts research based on job positions of IT professionals posted n web portals, designs a machine learning process with an unsupervised approach, extracts occupational profiles, designs a multidimensional model, applies k-means clustering when determining clusters of job positions by similarity, and reports the results obtained.

Keywords: machine learning process; clustering; k-means; social demand; IT professionals.

INTRODUCTION

Machine learning (ML) is a branch of artificial intelligence that uses scientific computing, mathematics and statistics through automated techniques to solve problems based on classification, regression and clustering. In recent years, the job description “data scientist”, whose functions include data cleaning, data transformation according to the context of the domain and the correct application of ML algorithms to obtain the most accurate learning model, has become popular. Clustering is an unsupervised ML technique based on the discovery of patterns or clusters of objects according to their geometric position in the n-dimensional vector space as explained by Sandhu and cited by Alloghani et al. (2020). Clustering quality depends on the complexity, dimensionality and granularity of the dataset, statistics and data distribution; it is applicable to untrained datasets, it is suitable in the exploratory stages of large volumes of information, and supervised ML techniques can be applied for information predictability purposes according to the context surrounding the area of interest (Perez, 2014; Swamynathan, 2017; Deshpande, 2018).

Social demand refers to the need for service and product of the professional training process, expressed by interest groups, aimed at contributing to national development, as established by the quality assurance policy of university higher education and the national licensing and accreditation models (Ministerio de Educación, 2015; Sistema Nacional de Evaluación, Acreditación y Certificación de la Calidad Educativa [SINEACE], 2016). A starting point for this research study was the idea of extracting social demand in a systematized manner from employment web portals, which have been used since the 1990s as digital spaces to which employers in the public and private sector regarded as representatives of interest groups turn to as they advertize their needs for required job profiles in these spaces. Employment is defined as a set of tasks and duties performed or intended to be

1 School of Graduate Studies - Universidad Nacional Federico Villareal (Lima, Peru). Currently, she is coordinator of the research group Ingeniería Web and associate professor at the School of Systems Engineering and Informatics of Universidad Nacional Mayor de San Marcos (Lima, Peru).
Orcid: https://orcid.org/0000-0002-2590-8387
E-mail: zmamanir@unmsm.edu.pe

performed by a person. Similarly, an occupation is a type of work performed in a job. A person may be associated with one or several jobs performed over time, which strengthens their resume (Organización Internacional del Trabajo [OIT], 2012).

An excerpt of the details of a job posting is shown in Table 1. Line 1 refers to the job title; line 2 specifies the job location; line 3 shows the requirements, which can be broken down into simple components such as education (line 4), work experience (line 5), training (line 6) and knowledge (line 7); line 8 lists the functions, which refer to the roles and/or responsibilities involved in the job position, the same that are made explicit in lines 9-11; lines 12-13 specify the training required; lines 14-15 refer to the skills that the job applicant must have; and lines 16-17 specify the type of contract offered by the employer.

The term information technology (IT) was first used in the 1958 Harvard Business Review to describe the "new technology" in business, associating it with the use of computer processing of information, mathematical programming for decision-making and simulation through computer programs, as noted by Leavitt and Whisler (1958), as cited in Rowe et al. (2011); in the last four years, the Information Technology Curricula 2017 proposed by the Association for Computing Machinery (ACM) and the Institute of Electrical and Electronics Engineers (IEEE) has redefined IT as "the study of systemic approaches to select, develop, apply, integrate, and administer secure computing technologies to enable users to accomplish their personal, organizational, and societal goals" (p. 18).

> The IT graduate is a collaborative problem solver, skilled practitioner, or applied research investigator who enjoys getting technology to work effectively and meet user needs in a variety of settings. IT graduates work collaboratively to integrate new technologies in the workplace and community and ensure a superior and productive experience for the user and all the organization's functions. In the corporate environment, IT graduates apply their understandings of system integration, development, and operation and deploy and manage IT services and platforms that meet the business goals and objectives of the organization. In the community, IT graduates use their expertise

Table 1. *Detail of a Job Posting.*

1.	**Tics Application Programmer Analyst CAS 023**
2.	**Lima**
3.	**REQUIREMENTS**:
4.	Bachelor's Degree in Systems Engineering, Software Engineering, Computer Engineering or related.
5.	Minimum of three (03) years of experience working in public or private organizations.
6.	Minimum of Java programming 24-hour course and at least Scrum, Kanban or Agile methodologies 24-hour courses.
7.	Knowledge of Java platform, javascript, json, HTML5, CSS, SOAP and REST services consumption, Angular, PL / SQL, microservices, agile development methodologies (Scrum) and the Peruvian Technical Standard: NTP 12207.
8.	**Job Functions**
9.	Perform adaptive and perfective maintenance to the existing systems in the institution, according to the functional and operational needs in order to maintain the operability of the information system
10.	Perform the implementation process in the production environment of the functionality, module or information system, so that the user areas have an application according to their needs
11.	Support and provide initial training in the implementation process of the functionality, module or information system developed, in order to ensure the correct operation of the systems
12.	**Requirements**
13.	Degree in Computer Engineering
14.	**Skills**
15.	Teamwork, Customer-oriented, Result-oriented
16.	**Contract type**
17.	Contract for work and services

Source: Excerpted from Google Jobs Search, Google (2021).

> in implementing a wide range of IT solutions to support community members' projects and activities. IT graduates are professionals prepared to perform duties in an ethical manner. They are familiar with the various laws and regulations that govern the development and operations of the IT platforms they maintain. IT graduates can explain and justify professional decisions in a language that both management and clients understand. They are aware of the budget implications of technological alternatives and can defend budgets properly. IT graduates have extensive practice with properly securing IT networks, applications, data centers, and online services. They seek secure technology solutions without unduly adversely affecting the ability of users to accomplish their goals. (Association for Computing Machinery & IEEE Computer Society, 2017, p. 19)

Related works include the research by Qin et al. (2018) who proposed a semantic model to improve person-job matching for online talent recruitment, for which the authors establish a semantic representation of job advertisements and candidates' resumes. In the experimental stage, they use the dataset of a Chinese technology company and several supervised ML techniques such as logistic regression, decision trees, random forests, and gradient boosting decision tree to assess the accuracy and efficiency of the results.

Similarly, Boselli et al. (2018) study the classification of online job advertisements by means of supervised ML, their contribution lies in the extraction of job advertisements from web portals. For this purpose, they apply webscraping, the dataset is trained by domain experts consigning the ISCO taxonomy to the profiles and generate machine learning models with linear support vector machine (SVM), RBF Kernel SVM, random forests (RF) and neural networks techniques. For the extraction of skills from job postings, they use n-gram text categorization, filter low significance n-grams, and involve domain experts to establish the ESCO taxonomy.

Lynch (2017) aims to solve an organizational problem concerning human resources personnel, who subjectively determine the job profiles, salary, level and responsibility of employees, based on the job description, which may result in bias and inconsistencies; therefore, his research focuses on the analysis of the predictability of job titles based on the job description. Lynch collects the job titles from a web page, transforms them using natural language processing (NLP) and obtains a dataset of keywords determined based on term frequency; supervised machine learning techniques such as SVM and RF are applied to the resulting model to predict the top 30 most frequently occurring job titles.

Marrara et al. (2017) propose an occupational recognition approach on the ISCO taxonomy based on the linguistic model; it describes a plausible improvement of the WoLMIS project. The experimental testing demonstrated its potential to identify potential new occupations from the analyzed job postings.

The research by Vinel et al. (2019) deals with the experimental comparison of unsupervised approaches for discovering specializations of professions within the job posting corpus. Various statistical methods of text vector representations are experimentally evaluated: TF-IDF, additive regularization of topic models (ARTM), neural language models based on distributional semantics (word2vec, fasttext) and deep contextualized word representation (ELMo and multilingual BERT). A Russian job vacancy dataset is used. and k-means clustering, affinity propagation, BIRCH, agglomerative clustering and HDBSCAN techniques are used. They conclude that as long as the number of clusters k-means to be obtained is specified in advance the best solution is obtained by ARTM, if not, word2vec is better.

In this context, the following objectives are established: 1) design an unsupervised machine learning process, 2) extract the social demand from job web portals using webscraping techniques, 3) pre-process the information using natural processing language techniques, 4) design a multidimensional model, 5) populate the multidimensional model, 6) apply the unsupervised machine learning technique k-means, and 7) evaluate the resulting machine learning model.

The main contribution of this research study consists of proposing and implementing an unsupervised machine learning process that defines a set of activities oriented to extract the social demand for IT professional jobs from job portals using automated techniques and to create a k-means machine learning model based on the recognition of job profiles; additionally, visualization perspectives are presented using business intelligence techniques. It is considered relevant and novel because its application would allow knowing, through automation, the social demand for professional careers related to Information Technologies, which could be applied to other disciplinary areas, being a topic of interest for the actors responsible for the management of academic programs for licensing and accreditation

purposes. It can also contribute to the updating of the Catálogo Nacional de Perfiles Ocupacionales [Peruvian Catalog of Occupational Profiles] (CNPO) and to further enrich the repository of qualifications of Peru (MNCP).

METHODOLOGY

According to Hernández et al. (2014), this is a qualitative research study, as it does not attempt to correlate variables; it uses the inductive method, as it is based on case studies; due to the time of application of the variables, it is cross-sectional, as the data will be collected at a single moment and time; and, due to the nature of the objectives, it is a descriptive, non-experimental and applied research, because it uses existing knowledge in order to find solutions to the problems raised.

The population is made up of the job positions of the groups of interest, these being employers representing the public and private sector, who use job portals to launch public job advertisements so as to carry out a transparent selection process and recruit the best professionals who meet the required criteria. The sampling technique to be used is non-probabilistic and intentional, as it is considered key in providing valuable information for the research, and includes job ads posted in the last two years. The sample was comprised of 8640 job ads published between February 2020 and February 2021 in the following web portals: i) Google Jobs Search (3200), ii) Freelancer (2096), iii) Buscojobs (1289), iv) Mipleo (724), v) Linkedin (457), vi) Indeed (420), vii) Computrabajo (379), viii) Convocatoriatrabajo (75).

Research starts with the design of a machine learning process with an unsupervised approach which subprocesses and activities have been customized, as shown in Figure 1, where the design formulated by Swamynathan (2017, p. 195) is adapted based on the stages of the traditional data mining processes Knowledge Discovery Databases (KDD) and its variant Cross-Industry Standard Process for Data Mining (CRISP-DM).

The process includes two sub-processes. The first involves extracting data related to job profiles of IT professionals posted on the main web platforms using webscraping techniques and storing it in a database in a custom designed schema. A Python program including customized functions was developed for each job portal, considering that each one contains its own implementation characteristics for modeling data in html, some of which are more complex than others. The procedure consisted of entering Google Chrome web inspector, specifically, the elements tab through which users can access and read the document object model (DOM) of the

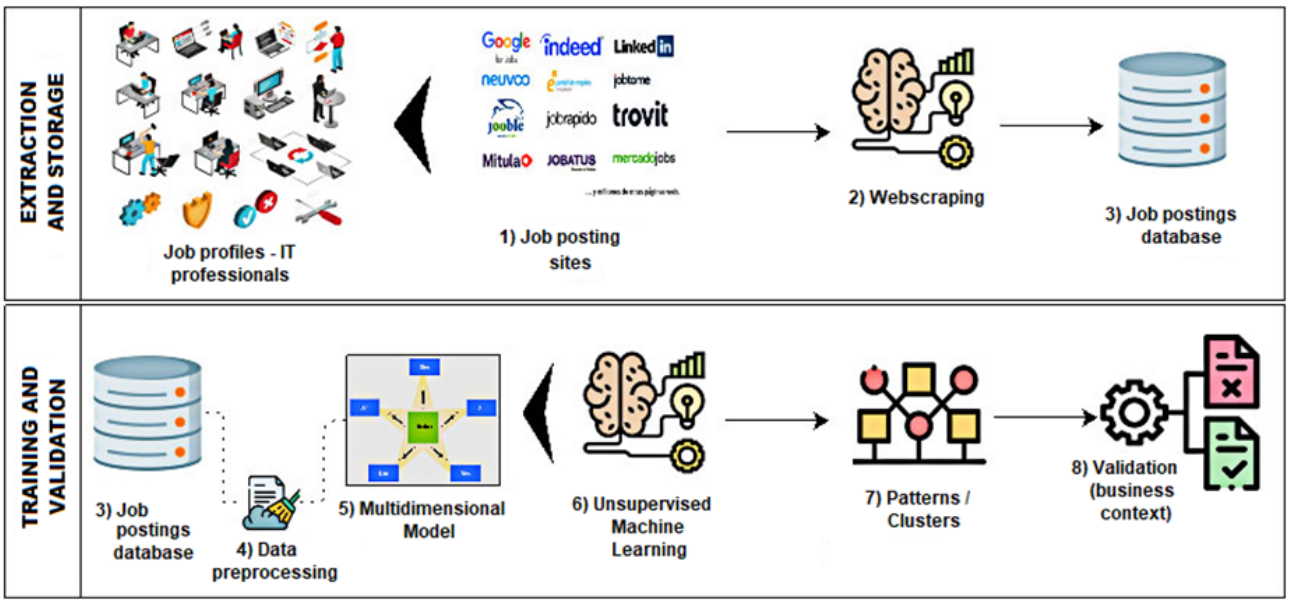

Figure 1. Unsupervised Machine Learning.

Source: Prepared by the author.

job advertisement, which structures the content of a document on the web, generally based on html tags of type list (li) or divisions (div), as indicated by MDN Web Docs (2005), whose contents were extracted with the support of the BeautifulSoup library.

Table 2 shows an excerpt of the program for the extraction of job postings in Python programming language. From lines 1 to 6, the libraries required by the program are imported; in line 7, function maledpeti_portal() is defined, containing the programming logic to access the database (line 9), obtain the keywords by calling the getWords() method (line 10). In line 11, an iterative process is performed according to the words retrieved in line 10, and for each word it will perform the following actions: initialize the load array with the values set in the constants WS_PORTAL_LABORAL, WS_PAGINAS, WS_PAGINA_INICIAL, WS_OFERTAS, WS_AREA, which contain information related to the job portal to process, the number of pages to read, the start page to read, the number of job offers or positions to view, as well as the search area, respectively, as a job portal contains information of job positions from different disciplinary areas. The search parameters are set in line 20; in line 21, the registrar_webscraping() method is called, which is defined in the controller class, resulting in database schema population: webscraping and job offer tables. To register the job description, the actions listed in line 23, 24 and 26 should be performed, depending on the type of job portal, considering that different portals describe a different html structure; this procedure is repeated for each keyword to be searched for and considered in the word array (line 10). The main method can be observed line 28, from which the function maledpeti_portal() is invoked using the portal to be processed as a parameter to populate the table oferta_detalle.

Table 2. *Excerpt of the Program for the Extraction of Job Postings.*

```
from configuration import *
import webscraping_computrabajo
import webscraping_indeed
from controller import Controller
import datetime
import webscraping_buscojobs
def maledpeti_portal(sitio):
controller = Controller()
con = connect_bd()
palabras = controller.getwords(con)
for filtro in palabras:
carga = {}
carga["pagina"] = sitio["WS_PORTAL_LABORAL"]
carga["cant_paginas"] = sitio["WS_PAGINAS"]
carga["pagina_inicial"] = sitio["WS_PAGINA_INICIAL"]
carga["cant_ofertas"] = sitio["WS_OFERTAS"]
carga["busqueda_area"] = sitio["WS_AREA"]
carga["busqueda"] = ""
carga["id_keyword"]=filtro[0]
set_url_busqueda(carga, sitio, filtro[1])
carga["id_carga"] = controller.registrar_webscraping(con, carga)
if sitio["WS_PORTAL_LABORAL"]=="computrabajo":
listaOferta = webscraping_computrabajo.scraping_ofertas(con, carga["url_principal"], carga["url_prefix"], carga["url_sufix"],carga["pagina_inicial"], carga["cant_paginas"], carga["cant_ofertas"],carga["id_carga"])
elif sitio["WS_PORTAL_LABORAL"]=="indeed":
listaOferta = webscraping_indeed.scraping_ofertas(con, carga["url_principal"], carga["url_prefix"], carga["url_sufix"],carga["pagina_inicial"], carga["cant_paginas"], carga["cant_ofertas"],carga["id_carga"])
elif sitio["WS_PORTAL_LABORAL"]=="buscojobs":
listaOferta = webscraping_buscojobs.scraping_ofertas(con, carga["url_principal"], carga["url_prefix"], carga["url_sufix"],carga["pagina_inicial"], carga["cant_paginas"], carga["cant_ofertas"],carga["id_carga"])
if __name__ == "__main__":
maledpeti_portal(COMPUTRABAJO)
maledpeti _portal(INDEED)
maledpeti _portal(BUSCOJOBS)
```

Source: Prepared by the author.

The second subprocess is the training and validation of the model, this is subdivided into six activities, which are detailed below.

Preprocessing. It consists in using text mining techniques for the segmentation of the job description into subunits that are stored in individual tuples in a database schema for their subsequent classification according to the type of dimension involved. As noted by Swamynathan (2017, pp. 70, 256), outliers and errors were removed to ensure an efficient analysis. It consisted of cleaning the text by converting text-type values to uppercase, removing checkmarks, strange characters, blank spaces, texts or words without meanings, and tokenizing the job posting. The latter consisted of breaking the job description into simple meaningful components to which the term "dimension" was assigned.

Table 3 shows an excerpt of the Python code used to remove noise. For instance, in line 17, the remove_tags_html() function is invoked to remove html tags found in the text string; in line 18, the incomplete_tags_html() function is called to remove incomplete html tags found in the job post description text; and, in lines 19 and 20, the remove_non_ascii() and remove_space() functions are cited to remove non-ascii codes and blank spaces or characters from the beginning and end of the string or paragraph of text corresponding to the job post description.

Design of the Multidimensional Model. The purpose of this activity was to organize the elements comprising a job position. Following Kimball and Ross (2002), the snowflake scheme characteristic of a high granularity data mart was used, consisting of a fact table "job positions" and 14 main dimensions such as category, job title, profile, web portal, employer, city, salary, period, function, knowledge, competencies, skills, qualifications, benefits and education, as can be seen in Figure 2.

Using a multidimensional scheme to model the information made it possible to appreciate the data from different perspectives, as can be seen in Figure 3, which shows the set of dashboards of the project created with the business intelligence tool Power BI 2.99.862.0. The information refers to job positions for IT professionals, organized by period, website, company, functions, competencies, skills, benefits and salary; it is also possible to incorporate other dimensions. Four main perspectives are observed, the first one displays the percentage of job positions per web page, where the highest percentage for 2021 corresponds to Google Jobs Search

Table 3. *Python Code for Noise Removal.*

```
from configuration import *
from controller import Controller
from preprocessing import PreProcessing
from dbconnection import Connection
import numpy as np
from bs4 import BeautifulSoup

if __name__ == "__main__":
controller = Controller()
preprocessing = PreProcessing()
con = connect_bd()
oferta_detalle = controller.dbofertadetalle
datos = oferta_detalle.select_ofertadetalle_dimension(con,3)
datos = np.array(datos)
i = 1
matrix = [fila[i] for fila in datos]

#normalizar data
matrix = preprocessing.remove_tags_html(matrix)
matrix = preprocessing.remove_incomplete_tags_html(matrix)
matrix = preprocessing.remove_non_ascii(matrix)
matrix = preprocessing.remove_space(matrix)
#modificar la columna de descripcion
for indice in range(0,len(matrix)):
datos[indice][1] = matrix[indice]
```

Source: Prepared by the author.

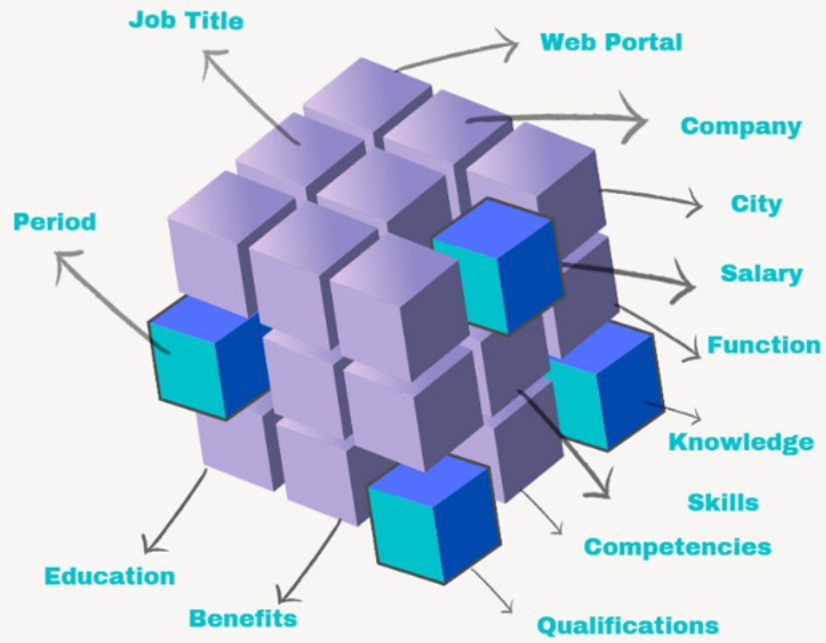

Figure 2. Multidimensional Model.

Source: Prepared by the author.

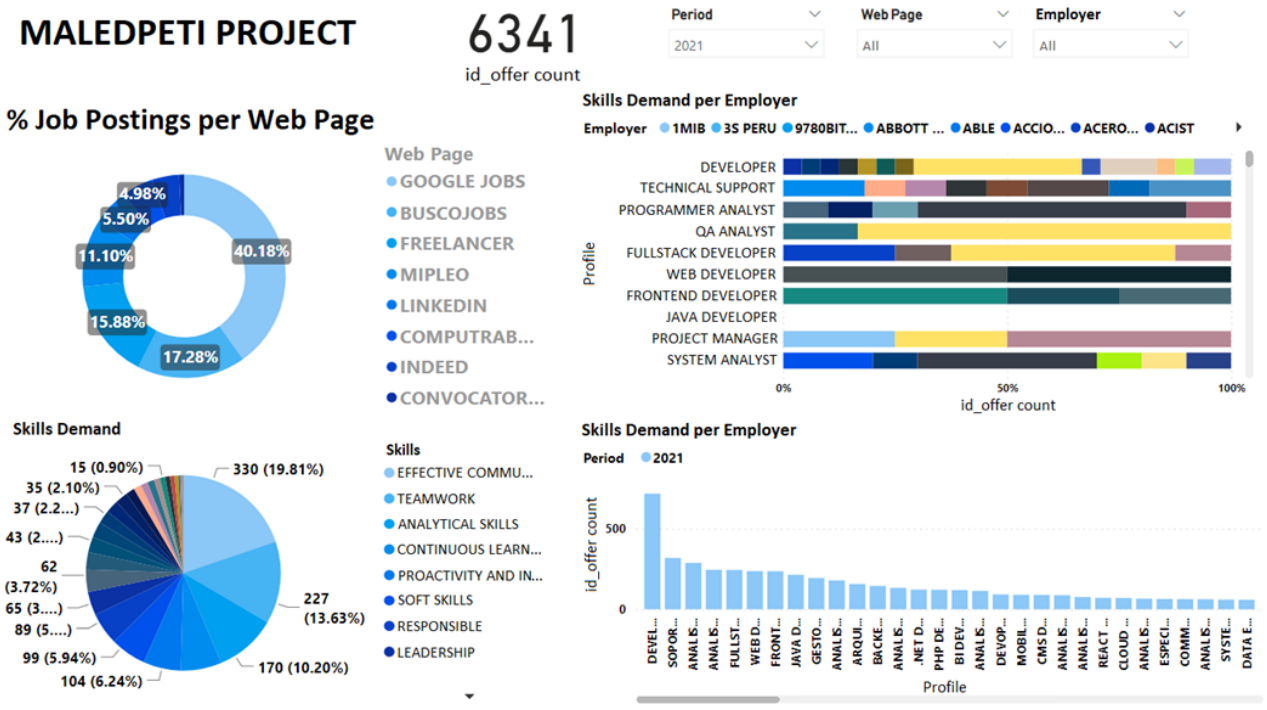

Figure 3. MALEDPETI Project Dashboards.

Source: Prepared by the author.

with 40.18%, while the other web pages exhibit a smaller percentage for the same period.

A pie chart shows the transversal, soft, social and organizational skills that employers usually require; it is evident that effective communication, teamwork, analytical skills, proactivity and initiative are the most in-demand skills in the job profiles of IT professionals. Additionally, it shows the main IT profiles required by employers such as banks, financial companies, mining companies, service companies and retailers; as for the public sector, ministries, municipalities, among other governmental institutions. The fourth dashboard shows the demand for job profiles by period, where "developer", "technical support", "programmer analyst", "quality analyst", "fullstack developer" were among the most in-demand in 2021.

Unsupervised Machine Learning. At this stage, clustering, an unsupervised technique based on the determination of clusters grouped by similarity of data points with respect to their geometric position in the vector space, is applied; it is a descriptive and classification technique independent of any formal model, which does not assume the existence of dependent or independent variables, and does not require trained datasets for its analysis; the models are created automatically based on data recognition (Perez, 2014; Swamynathan, 2017; Deshpande, 2018). The logic of the k-means algorithm is defined in Sierra (2006) as shown in Figure 4:

Additional to the proposals of Sierra (2006) and Swamynathan (2017), there are several variants of the k-means algorithm in the literature, focused on achieving good quality of the clustering results and avoid depending on the number of clusters to obtain. The non-convergence in results can be mentioned among the incidents reported in the application of the algorithm, which may be due to the number of clusters *k* chosen or to the absence of cluster structure in the data, according to Sierra (2006, p. 290). The k-means++ algorithm used in this research was proposed by Arthur and Vassilvitskii (2007), and it is a way of initializing k-means based on the determination of random centroids with very specific probabilities that achieves substantial improvements in terms of precision and speed with respect to the traditional Lloyd's k-means method. The kmeans++ algorithm requires to specify the value of *k* clusters to be obtained as an input parameter. Therefore, following Nerurkar et al. (2018), the Elbow method was used to estimate the ideal number of *k* clusters as a function of the inflection point in the graph; the objective function of the algorithm is to minimize the sum of squared distances in the *k* clusters. The formula can be seen in expression 1:

i. From among the n cases choose k which we will call seeds and set c_j , $j = 1, \cdots k$, Each seed c_j will represent cluster $C_j (j = 1, \cdots k)$.

ii. Assign case i to cluster C_j when $d(x_i, c_j) = \text{mín}_{l=1,\ldots,k,}\, d(x_i, c_l)$. That is, each case is assigned to the cluster with the closest seed point.

iii. Steps 1 and 2 provide an initial partitioning of the cases.

iv. Calculate the improvement that would occur in the chosen criterion (minimize $tr(W)$, minimize $det(W)$, etc…) by assigning a case to another cluster.

v. Make the change that leads to the greatest improvement in the clustering criterion.

vi. Repeat steps 3 and 4 until the chosen criterion stops improving.

Figure 4. K-means Algorithm.

Note. The figure displays the logical sequence of the k-means clustering algorithm.

Source: Taken from *Aprendizaje Automático conceptos básicos y avanzados, Aspectos prácticos utilizando el software WEKA* (p. 290), by B. Sierra, 2006, Pearson Prentice Hall.

$$\text{SSE} = \sum_{k=1}^{k} \sum_{(x_i \in k)} \|x_i - c_k\|^2, where\ c_k = centroid\ of\ the\ cluster. \quad (1)$$

Source: Taken from *Empirical Analysis of Data Clustering Algorithms* (p. 2), by Nerurkar et al. (2018), Procedia Computer Science 125(1).

Using the k-means technique with software Weka 3.8.5, a cluster-based model was determined; the k-means algorithm implemented by Weka is the one proposed by Arthur and Vassilvitskii (2007). Several functions are available from the interface for the calculation of distances: Chebyshev, Euclidean, Filtered, Manhattan and Minkowski. Another relevant aspect is the initialization method: Random, k-means++, Canopy and Farthest first, which are parameters required by the algorithm, as well as the number of clusters to determine and the maximum number of iterations.

Line 1 of Table 4 shows the algorithm execution parameters in Weka with a fourteen-dimensional dataset, 15 clusters, k-means++ as initialization method and euclidean as distance function. The result yielded the following metrics: 7145 instances evaluated, 6 iterations performed and a sum of squared errors of 34 696.82. Table 5 shows that cluster #5 is the most significant cluster as it groups the highest number of instances. It shows that 17% of the job positions for IT professionals correspond to 2020 and to the job profile "fullstack developer" posted in "Google Jobs Search" to work in a company in the city of "Lima-Peru", the salary is "not specified", the main function to be performed is "development and implementation of software projects", the applicant is required to demonstrate "knowledge of database", the main competency is "development in Java programming language", "effective communication" is one of the important soft skills for the job, the applicant must be "certified in Scrum", the company offers "job stability", and the applicants must be "graduates or Bachelors of Systems Engineering, Computer Science or related".

Considering a reduction of the dimensionality of the dataset to six, the resulting metrics are shown in line 2 of Table 4, where it can be observed that the algorithm performed 4 iterations with a sum of squared error of 16 027.00, much lower with respect to the 14-dimensional dataset. Table 6 shows cluster #3 with 2318 instances (33%) as the most representative and reveals that job profile "developer" is the most requested by employers, the main function is "development and implementation of software projects", "database knowledge" is required, "competence in Java programming language" must be demonstrated, "effective communication" is one

Table 4. *Results of K-means Clustering with Weka.*

No.	Technique	Instances	No. of Dim.	No. of Clusters	Init. Method	No. of Iter.	SE^2 (Intra-Cluster)
1	Kmeans/Weka	7 145.00	14	15	k-means++	6	34 696.82
2	Kmeans/Weka	7 006.00	6	15	k-means++	4	16 027.00

Source: Prepared by the author.

Table 5. *Cluster #5 14-Dimensional Dataset.*

Cluster #5 Instances: 1239 (17%)
Category: DEVELOPER
Job Profile: FULLSTACK DEVELOPER
Web page: GOOGLE JOBS SEARCH
Company: NOT SPECIFIED
City: Lima
Salary: NOT SPECIFIED
Period: 2020
Functions: DEVELOPMENT AND IMPLEMENTATION OF SOFTWARE PROJECTS
Knowledge: DATABASE KNOWLEDGE
Competencies: JAVA
Skills: EFFECTIVE COMMUNICATION
Qualifications: SCRUM CERTIFICATION
Benefits: JOB STABILITY
Education: GRADUATE OR BACHELOR'S DEGREE IN SYSTEMS ENGINEERING, COMPUTER SCIENCE OR RELATED FIELDS

Source: Prepared by the author.

Table 6. *Cluster #3 6-Dimensional Dataset*

Cluster #3 Instances: 2318 (33%)
Job Profile: DEVELOPER
Functions: DEVELOPMENT AND IMPLEMENTATION OF SOFTWARE PROJECTS
Knowledge: DATABASE KNOWLEDGE
Competencies: JAVA
Skills: EFFECTIVE COMMUNICATION
Benefts: JOB STABILITY

Source: Prepared by the author.

of the most requested soft skills, and the greatest benefit offered by the employer is “job stability”.

RESULTS AND DISCUSSION

The design of an unsupervised ML process made it possible to clearly establish the steps follow to determine the social demand for IT professional jobs. The proposed process involved two sub-processes: the first one comprised the extraction and storage of data concerning the employment profiles of IT professionals, obtained from the main web platforms using Webscraping techniques, and the second sub-process involved clustering and model validation activities.

Upon completion of the experimental phase of the k-means unsupervised ML technique with Weka, the metrics obtained were shown in Table 4. Considering the 14-dimensional dataset, the “sum of squared errors” metric was higher than the 6-dimensional one; the impact on the number of iterations required to form the 15 clusters is also shown, which was reduced from six to four iterations.

Using a multidimensional model to model the information of the job positions made it possible to obtain different perspectives on the behavior of the information, such as the functions required for the job profiles, some of which are shown in Figure 5. The same function is required for several job profiles, each color of the bar represents a profile, for example: the function “project management” is one of the tasks that would be performed by the profiles “web developer”, “system administrator”, “project manager”, “developer”, “CMS developer”, “software architect”, “IT security analyst”, “QA analyst”, “Cloud analyst” and “network and communications administrator”. Similarly, the other functions outlined in Figure 5 can be analyzed.

Another important factor in a job posting is the benefits offered by employers, the highest percentage of which corresponds to the legal rights according to the type of position. Based on the results, the most frequently offered benefit is “job stability”, followed by “career path”, “corporate benefits”, “health insurance” and “continous learning”, while very few employers offer “profit sharing”.

A high variability of salaries offered by private sector companies and, to a lesser extent, by public organizations has been determined. A large percentage of companies do not specify the salary in the job offer, they tend to state “not specified”, “commensurate with the market rate”, “commensurate with responsibility”, “commensurate with experience and knowledge”, “commensurate with specific project” and “subject to evaluation”. Other employers did state the salaries offered, as was the case for “developer” and related job profiles, for which the salaries offered range between <3000; 6000> soles.

The competencies that the applicant for an IT job position must possess may not be exclusive to a particular profile. An example of this is the technical competency “designing the software architecture”, which is more relevant for the “developer” category, but is also required for “architect” and “manager” categories, and is less so for “systems analyst” category. The other competencies can be analyzed under the same logic.

Also, the “data engineer” profile requires demonstrating compliance with technical competencies such as AWS, database, big data, cloud computing, data science, data warehouse, software development, development platform, ETL and process management. Similarly, the competencies required in each job profile can be analyzed systematically.

Soft skills are other elements of information of high value in a job position, among the most demanded transversal skills are “effective communication”, “teamwork” and “analytical skills”; “resilience” is a skill that is being increasingly demanded given the current pandemic context.

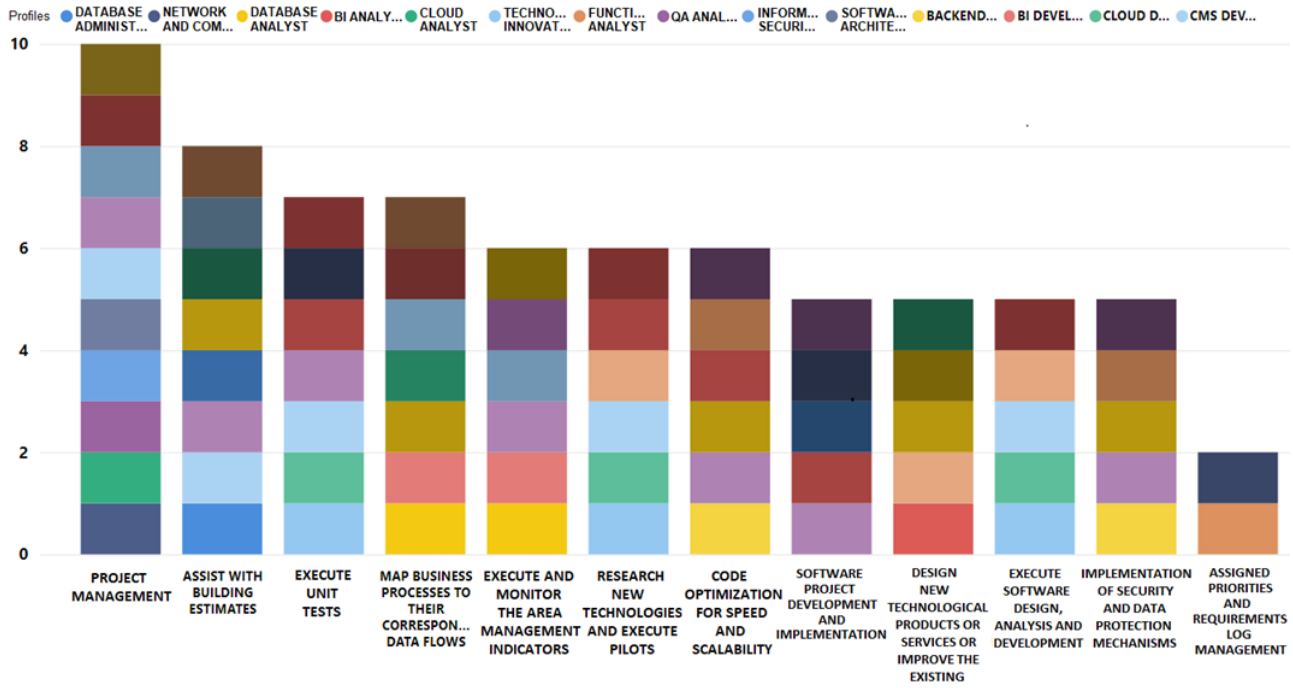

Figure 5. Transversal Functions to Job Profile.

Source: Prepared by the author.

CONCLUSIONS

1. This research study proposes an unsupervised ML process to determine the social demand for IT professional jobs. The proposed process involved two sub-processes: the first one comprises the extraction and storage of data concerning the employment profiles of IT professionals, obtained from the main web platforms using webscraping techniques, and the second sub-process involves clustering and model validation activities.

2. A multidimensional model consisting of a fact table “job positions” and 14 main dimensions such as category, job title, profile, web portal, employer, city, salary, period, function, knowledge, competencies, skills, qualifications and benefits was designed. Supported by a business intelligence tool, it allowed to determine the behavior of social demand through different perspectives, according to the dimensions identified for this purpose.

3. Unsupervised ML was applied using the k-means clustering technique to determine the behavior of the social demand, based on the clusters generated by similarity of points with respect to their geometric position in the vector space, creating models automatically from the recognition of the data. The parameters used in the execution of the algorithm, the number of iterations determined and the intra-cluster metric are shown in Table 4.

4. The characteristics of the identified patterns, considering 14 and 6 dimensions respectively, are shown in Tables 5 and 6, and their significance is detailed in the results and discussion section.

5. This proposal can be used in three scenarios: i) to systematize social demand, relevant to decision makers in the management of academic programs for licensing and accreditation purposes, ii) to update the Peruvian Catalog of Occupational Profiles (CNPO), according to the social demand for IT professions and also to other areas of knowledge, and iii) to systematically populate the repository of qualifications of Peru (MNCP).

REFERENCES

[1] Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., y Aljaaf, A. (2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. En M. W. Berry, A. Mohamed, Nee Wah P. (Eds.), *Supervised and Unsupervised Learning forp Data Science* (pp. 3-21). Cham, Suiza: Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-22475-2

[2] Arthur, D., y Vassilvitskii, S. (2007, 7-9 de enero). k-means++: the advantages of careful seeding. *Eeighteenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 1027-1035). New Orleans, LA, EE. UU. https://dl.acm.org/doi/10.5555/1283383.1283494

[3] Association for Computing Machinery, y IEEE Computer Society. (2017). *Information Technology Curricula 2017: Curriculum Guidelines for Baccalaureate Degree Programs in Information Technology.* https://doi.org/10.1145/3173161

[4] Boselli, R., Cesarini, M., Mercorio, F., y Mezzanzanica, M. (2018). Classifying online job advertisements through machine learning. *Future Generation Computer Systems*, *86*, 319-328. https://doi.org/10.1016/j.future.2018.03.035

[5] Deshpande, M. (2018). *Machine Learning for Human Beings. Build Machine Learning Algorithms with Python.* Zenva Pty Ltd.

[6] Google. (2021). *Google Search Jobs*. https://www.google.com/search

[7] Hernández, R., Fernández Collado, C., y Baptista, P. (2014). *Metodología de la Investigación* (6ª ed.). México D. F., México: McGraw-Hill/Interamericana Editores.

[8] Kimball, R., y Ross, M. (2002). *The data warehouse toolkit : the complete guide to dimensional modeling* (2ª ed.). Inidianapolis, IN, EE. UU.: John Wiley & Sons.

[9] Leavitt, H., y Whisler, T. (1958). Management in the 1980's. *Harvard Business Review, 36*, 41-48.

[10] Lynch, J. (2017). *An Analysis of Predicting Job Titles Using Job Descriptions.* (Dissertation for M.Sc. in Computing). Technological University Dublin, Dublin. https://arrow.dit.ie/cgi/viewcontent.cgi?article=1125&context=scschcomdis

[11] Marrara, S., Pasi, G., Viviani, M., Cesarini, M., Mercorio, F., Mezzanzanica, y Pappagallo, M. (2017). *A language modelling approach for discovering novel labour market occupations from the web.* WI '17: International Conference on Web Intelligence. Leipzig, Alemania. https://doi.org/10.1145/3106426.3109035

[12] Ministerio de Educación. (2015). *Política de Aseguramiento de la Calidad.* http://www.minedu.gob.pe/reforma-universitaria/pdf/politica.pdf

[13] MDN Web Docs. (2005). *Introducción al DOM*. https://developer.mozilla.org/en-US/docs/Web/API/Document_Object_Model/Introduction

[14] Nerurkar, P., Shirke, A., Chandane, M., y Bhirud, S. (2018). Empirical Analysis of Data Clustering Algorithms. *Procedia Computer Science*, *125*, 770-779. https://doi.org/10.1016/j.procs.2017.12.099

[15] Organización Internacional del Trabajo. (2012). *Clasificación Internacional Uniforme de Ocupaciones*. https://www.ilo.org/public/spanish/bureau/stat/isco/isco08/index.htm

[16] Perez, C. (2014). *Técnicas de minería de datos e inteligencia de negocios: IBM SPSS Modeler.* Madrid, España: Ibergarceta Publicaciones, S.L.

[17] Qin, C., Zhu, H., Xu, T., Zhu, C., Jiang, L., Chen, E., y Xiong, H. (2018). *Enhancing Person-Job Fit for Talent Recruitment: An Ability-aware Neural Network Approach*. SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. Nueva York, NY, EE. UU. https://doi.org/10.1145/3209978.3210025

[18] Rowe, D., Lunt, B., y Helps, R. (2011, 20-22 de octubre). *An Assessment Framework for Identifying Information Technology Programs* [Conferencia]. 2011 ACM Conference on Information Technology Education. Nueva York, NY, EE. UU. https://doi.org/10.1145/2047594.2047630

[19] Sierra, B. (2006). *Aprendizaje Automático. Conceptos básicos y avanzados. Aspectos prácticos utilizando el software WEKA.* Madrid, España: Pearson Prentice Hall.

[20] Sistema Nacional de Evaluación, Acreditación y Certificación de la Calidad Educativa. (2016). *Modelo de Acreditacion para Programas de Estudios de Educacion Superior Universitaria.* https://hdl.handle.net/20.500.12982/4086

[21] Swamynathan, M. (2017). *Mastering Machine Learning with Python in Six Steps, A Practical Implementation Guide to Predictive Data Analytics Using Python.* Berkeley, CA, EE. UU.: Apress Media. https://doi.org/10.1007/978-1-4842-2866-1

[22] Vinel, M., Ryazanov, I., Botov, D., y Nikolaev, I. (2019). *Experimental Comparison of Unsupervised Approaches in the Task of Separating Specializations Within Professions in Job Vacancies*. 8th Conference, AINL 2019 (pp. 99-112). Tartu, Estonia: Springer Nature Switzerland AG 2019. https://doi.org/10.1007/978-3-030-34518-1_7