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

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

Proceso de machine learning para determinar la demanda social de puestos de empleo de profesionales de TI

ZORAIDA MAMANI RODRIGUEZ [1]

RECIBIDO: 06/12/2021 ACEPTADO: 12/01/2022 PUBLICADO: 31/12/2022

RESUMEN

El *machine learning* es una rama de la inteligencia artificial que utiliza la computación científica, las matemáticas y la estadística a través de técnicas automatizadas para resolver problemas basados en clasificación, regresión y *clustering*. La demanda social refiere a la necesidad de servicio y producto del proceso de formación profesional, que expresan los grupos de interés, orientada a contribuir al desarrollo nacional, tal como lo establecen la política de aseguramiento de la calidad de la educación superior univeristaria y los modelos de licenciamiento y acreditación nacional. En ese contexto el presente trabajo realiza una investigación a partir de los puestos de empleo de profesionales de TI publicados en los portales web, diseña un proceso de *machine learning* con enfoque no supervisado, extrae los perfiles ocupacionales, diseña un modelo multidimensional, aplica *clustering k-means* en la determinación de conglomerados de los puestos de empleo por similitud y expone los resultados obtenidos.

Palabras clave: proceso machine learning; clustering; k-means; demanda social; profesionales de TI.

INTRODUCCIÓN

El *machine learning* (ML) es una rama de la inteligencia artificial que utiliza la computación científica, las matemáticas y la estadística a través de técnicas automatizadas para resolver problemas basados en clasificación, regresión y *clustering*. En los últimos años se ha popularizado el perfil «científico de datos», cuyas funciones comprenden la limpieza, transformación de los datos de acuerdo con el contexto del dominio y la aplicación correcta de algoritmos ML con la finalidad de obtener el modelo de aprendizaje más preciso. El *clustering* es una técnica ML no supervisada basada en el descubrimiento de patrones o conglomerados de objetos según su posición geométrica en el espacio vectorial *n* dimensional según lo explica Sandhu citado por Alloghani et al. (2020), la calidad del agrupamiento depende de la complejidad, dimensionalidad y granularidad del *dataset*, de las estadísticas y de la distribución de los datos; esta técnica es aplicable a *datasets* no entrenados, adecuados en las etapas exploratorias de grandes volúmenes de información, complementariamente se puede aplicar técnicas ML supervisadas con fines de predictibilidad en la información según el contexto del negocio de interés (Perez, 2014; Swamynathan, 2017; Deshpande, 2018).

La demanda social refiere a la necesidad de servicio y producto del proceso de formación profesional, expresada por los grupos de interés, orientada a contribuir al desarrollo nacional, tal como lo establecen la política de aseguramiento de la calidad de la educación superior universitaria y los modelos de licenciamiento y acreditación nacional (Ministerio de Educación, 2015; Sistema Nacional de Evaluación, Acreditación y Certificación de la Calidad Educativa [SINEACE], 2016). La presente investigación parte de la idea de extraer la demanda social de manera sistematizada desde los portales web de empleo, que desde la década de los noventa se utiliza como un espacio digital al que

1 Escuela Universitaria de Posgrado - Universidad Nacional Federico Villareal (Lima, Perú). Actualmente, es coordinadora del grupo de investigacion Ingeniería Web y docente asociada de la Facultad de Ingeniería de Sistemas e Informática de la Universidad Nacional Mayor de San Marcos (Lima, Perú).
Orcid: https://orcid.org/0000-0002-2590-8387
E-mail: zmamanir@unmsm.edu.pe

recurren los empleadores del sector gubernamental y privado considerados como representantes de los grupos de interés en vista de que publicitan en estos espacios sus necesidades de perfiles laborales requeridos. El empleo se define como un conjunto de tareas y deberes realizados o destinados a ser realizados por una persona. Asimismo, una ocupación es un tipo de trabajo realizado en un empleo. Una persona puede estar asociada con uno o varios puestos de empleo desempeñados en el tiempo, lo que afianza su hoja de vida (Organización Internacional del Trabajo [OIT], 2012).

En la Tabla 1 se expone un extracto del detalle de un puesto de empleo, en donde se puede apreciar que la línea 1 refiere al título del empleo; la línea 2 precisa el lugar del empleo; la línea 3 muestra los requisitos, los cuales se pueden desagregar en componentes simples tales como formación (línea 4), experiencia laboral (línea 5), capacitaciones (línea 6) y conocimientos (línea 7); la línea 8 etiqueta las funciones, las cuales refieren a los roles y/o responsabilidades que implican al puesto laboral, las mismas que se hacen explícitas en las líneas 9-11; las líneas 12 y 13 precisan la formación requerida, las líneas 14 y 15 hacen referencia a las habilidades que debe tener el postulante al puesto laboral; y en las líneas 16 y 17 se precisa el tipo de contrato ofrecido por el empleador.

El término tecnología de la información (TI) se utilizó por primera vez en la Revista de Negocios de Harvard de 1958 para describir la «nueva tecnología» en los negocios, asociándolo al uso de procesamiento informático de la información, la programación matemática para la toma de decisiones y la simulación a través de programas informáticos, así lo señalan Leavitt y Whisler (1958), como se citó en Rowe et al. (2011); en los últimos cuatro años en el Currículo en Tecnología de Información 2017 propuesto por la Association for Computing Machinery (ACM) y el Institute of Electrical and Electronics Engineers (IEEE) se redefine TI como el estudio de enfoques sistémicos para seleccionar, desarrollar, aplicar, integrar y administrar tecnologías informáticas seguras para permitir a los usuarios lograr sus objetivos personales, organizativos y sociales (Association for Computing Machinery y IEEE Computer Society, 2017).

El profesional de TI es un solucionador de problemas colaborativo e investigador calificado que

Tabla 1. *Detalle de un puesto de empleo.*

1.	**Analista Programador de Aplicaciones Tics CAS 023**
2.	Lima
3.	**REQUISITOS**:
4.	Bachiller Universitario en Ingeniería de Sistemas, Ingeniería de Software, Ingeniería de Computación, Ingeniería Informática o afines por la formación.
5.	No menor de tres (03) años de experiencia general en entidades públicas o privadas.
6.	Curso no menor a 24 horas de Java y, curso no menor a 24 horas de Scrum o Kanban o metodologías ágiles.
7.	Conocimientos en la plataforma Java, javascript, json, HTML5, CSS, consumo de servicios SOAP y REST, Angular, PL / SQL, microservicios, metodologías ágiles de desarrollo (Scrum) y la Norma Técnica Peruana: NTP 12207.
8.	**Funciones**
9.	Realizar el mantenimiento adaptativo y perfectivo a los sistemas existentes en la institución, de acuerdo con las necesidades funcionales y operativas con el fin de mantener la operatividad de los sistemas de información.
10.	Realizar el proceso de implementación, en el entorno de producción de la funcionalidad, módulo o sistema informático, con el fin de que las áreas usuarias tengan un aplicativo acorde a sus necesidades.
11.	Realizar el acompañamiento y la capacitación inicial en el proceso de implementación de la funcionalidad, módulo o sistema informático desarrollado, con el fin de garantizar el correcto funcionamiento de los sistemas.
12.	**Requirements**
13.	Grado en Ingeniería Informática
14.	**Skills**
15.	Teamwork, Customer-oriented, Results-oriented
16.	**Contract type**
17.	Contrato por obra y servicio

Fuente: Tomado de Google Search Job, Google (2021).

disfruta haciendo que la tecnología funcione de manera efectiva y satisfaga las necesidades de los usuarios en una variedad de entornos; trabaja en colaboración para integrar nuevas tecnologías en el entorno de trabajo, la comunidad y garantizar una experiencia superior y productiva para el usuario y todos los procesos de la organización. En el entorno corporativo, aplica sus conocimientos sobre integración, desarrollo y operación de sistemas e implementa y administra servicios y plataformas de TI que cumplen con las metas y objetivos comerciales de la organización. En la comunidad, los profesionales de TI utilizan su experiencia en la implementación de una amplia gama de soluciones de TI para apoyar los proyectos y actividades de los miembros de la comunidad. Los profesionales de TI están preparados para realizar tareas de manera ética, están familiarizados con estándares nacionales e internacionales que rigen el desarrollo y las operaciones de las plataformas de TI que mantienen; asimismo, pueden explicar y justificar las decisiones profesionales en un lenguaje que la gerencia, los usuarios o los clientes entiendan. Conocen las implicaciones presupuestarias de las alternativas tecnológicas y pueden defender los presupuestos adecuadamente. Tienen una amplia práctica en asegurar adecuadamente las redes de TI, las aplicaciones, los centros de datos y servicios en línea. Buscan soluciones tecnológicas seguras sin afectar indebidamente la capacidad de los usuarios para lograr sus objetivos (Association for Computing Machinery y IEEE Computer Society, 2017, p. 19).

Entre los trabajos relacionados se puede mencionar el trabajo de Qin et al. (2018) quienes propusieron un modelo semántico para mejorar la adecuación persona-trabajo para el reclutamiento de talentos en línea, para lo cual los autores establecen una representación semántica de los anuncios de empleo y las hojas de vida de los candidatos, en la parte experimental, utilizan el *dataset* de una compañía tecnológica de China y varias técnicas de *machine learning* supervisado como regresión logística, árboles de decisiones, bosques aleatorios y *gradient boosting decision tree* para evaluar la precisión y eficiencia de los resultados.

Asimismo, se cuenta con la investigación de Boselli et al. (2018), quienes se centran en la clasificación de ofertas de empleo en línea a través del aprendizaje automático supervisado, su contribución se delimita en la extracción de los anuncios de empleo de los portales web, aplican *webscraping*, el *dataset* es entrenado por expertos del dominio consignando los clasificadores ISCO para los perfiles y genera modelos de *machine learning* con las técnicas de máquina de soporte vectorial (SVM) lineal, SVM RBF Kernel, bosques aleatorios y redes neuronales. Para la extracción de habilidades desde las ofertas de empleo, utilizan el clasificador de texto n-gram, se depuran los n-grams con baja significancia, participan expertos del dominio para establecer la clasificación de habilidades de ESCO.

La propuesta de Lynch (2017) está orientada a resolver un problema organizacional de recursos humanos, quienes determinan de manera subjetiva los perfiles de empleo, salario, nivel y responsabilidad de los empleados, basándose en el detalle del puesto, con lo que se generan sesgos e inconsistencias, es así que su investigación se enfoca en el análisis de la predictibilidad de los títulos de los puestos de empleo a partir del detalle del puesto. Obtiene los puestos de empleo de una página web, aplica varias transformaciones con el uso de lenguaje de procesamiento natural (LPN) y obtiene un *dataset* de palabras clave determinadas a partir de la frecuencia de los términos en la información; al modelo resultante, le aplica técnicas de *machine learning* supervisado como máquina de soporte vectorial y bosques aleatorios para predecir los treinta puestos de empleo más frecuentes.

Marrara et al. (2017) proponen un enfoque de reconocimiento de ocupaciones sobre la taxonomía ISCO basado en el modelo lingüístico, el enfoque describe una posible mejora del proyecto WoLMIS. La evaluación experimental demostró el potencial del enfoque para identificar posibles nuevas profesiones a partir de las ofertas de trabajo analizadas.

El trabajo de Vinel et al. (2019) trata sobre la comparación experimental de enfoques no supervisados para descubrir especializaciones de las profesiones que se ubican en el cuerpo de las vacantes laborales, evalúa experimentalmente varios métodos estadísticos de representaciones de vectores de texto: TF-IDF, modelado probabilístico de temas (ARTM), modelos de lenguaje neuronal basados en semántica distribucional (word2vec, fasttext) y representación profunda de palabras contextualizadas (ELMo y BERT multilingüe), utiliza un *dataset* de puestos de empleo en ruso y técnicas de *clustering k-means*, propagación por afinidad, BIRCH, agrupación aglomerativa y HDBSCAN; concluyen que la mejor solución fue *k-means* con ARTM siempre que se especifique el número de *clusters* por obtener con antelación, de lo contrario word2vec resulta mejor.

Es en ese contexto que la presente investigación establece como objetivos: 1) diseñar un proceso de *machine learning* no supervisado, 2) extraer la demanda social desde los portales de empleo utilizando técnicas de *webscraping*, 3) realizar un pre-procesamiento de la información mediante el uso de técnicas de lenguaje de procesamiento natural, 4) diseñar un modelo multidimensional, 5) poblar el modelo multidimensional, 6) aplicar la técnica de *machine learning* no supervisada *k-means* y 7) evaluar el modelo de *machine learning* resultante.

El principal aporte de esta investigación se centra en proponer e implementar un proceso de *machine learning* no supervisado que define un conjunto de actividades orientadas a extraer con técnicas automatizadas la demanda social de puestos de empleo de profesionales de TI desde los portales de empleo y crear un modelo de *machine learning k-means* a partir del reconocimiento de los perfiles de empleo; complementariamente, se exponen perspectivas de visualización mediante técnicas de inteligencia de negocios. Se considera relevante y novedosa porque su aplicación permitiría conocer, de manera automatizada, la demanda social de carreras profesionales relacionadas a Tecnologías de la Información, que podría aplicarse a otras áreas disciplinarias, por tratarse de un tema de interés para los actores responsables de la gestión de programas académicos con fines de licenciamiento y acreditación. Asimismo, puede contribuir a la actualización del Catálogo Nacional de Perfiles Ocupacionales y al poblamiento del repositorio de cualificaciones del Perú (MNCP).

METODOLOGÍA

Considerando lo señalado por Hernández et al. (2014), la presente investigación tiene un enfoque cualitativo, pues no busca correlacionar variables; utiliza el método inductivo, pues se basa en estudio de casos; por el tiempo de aplicación de las variables, es transversal, pues los datos serán recolectados en un único momento y tiempo; y, por la naturaleza de los objetivos, es una investigación descriptiva, no experimental y es aplicada, debido a que se está haciendo uso de conocimientos existentes para encontrar soluciones a los problemas planteados.

La población está conformada por los puestos de empleo de los grupos de interés, estos son los empleadores que representan al sector público y privado, quienes utilizan los portales de empleo para realizar convocatorias públicas con la finalidad de llevar a cabo un proceso de selección transparente y reclutar a los mejores profesionales que cumplan con el perfil requerido. La técnica de muestreo a utilizar es no probabilística e intencional, por considerarse clave en el suministro de información de valor para la investigación, y comprende los anuncios de empleo registrados en los dos últimos años. Se utilizó una muestra n de 8640 anuncios de empleo publicados entre febrero de 2020 y febrero de 2021 en los siguientes portales web: 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).

La propuesta de investigacion inicia con el diseño de un proceso de *machine learning* con enfoque no supervisado cuyos subprocesos y actividades han sido personalizados de acuerdo a la presente investigación, tal como se aprecia en la Figura 1, en donde se adapta el diseño formulado por Swamynathan (2017, p. 195) con base en las etapas de los tradicionales procesos de minería de datos «Knowledge Discovery Databases» (KDD) y su variante «Cross-Industry Standard Process for Data Mining» (CRISP-DM).

El proceso contempla dos subprocesos. El primero comprende la extracción de la información concerniente a los perfiles de empleo de profesionales de TI publicados en las principales plataformas web mediante las técnicas de *webscraping* y su almacenamiento en una base de datos en un esquema diseñado para tal fin. Se desarrolló un programa en Python que incluía funciones personalizadas para cada portal de empleo, dado que cada uno contiene sus propias características de implementación para modelar la información en html, algunas más complejas que otras. En esencia, el procedimiento consistió en ingresar a la página de inspección del navegador Google Chrome, específicamente, al panel de elementos a través del cual se puede acceder y leer el modelo de objetos del documento (DOM) del anuncio de empleo, el cual estructura el contenido de un documento en la web, basado generalmente en etiquetas html de tipo lista (li) o divisiones (div), según lo indica MDN Web Docs (2005), cuyos contenidos fueron extraídos con el apoyo de la librería BeautifulSoup.

En la Tabla 2, se muestra un extracto del programa para la extracción de puestos de empleo en lenguaje de programación Python. En las líneas 1 a 6, se importan las librerías requeridas por el programa; en la línea 7, se define la función maledpeti_portal(), esta función contiene la lógica de programación para conectarse a la base de datos (línea 9), obtener los keywords o palabras clave invocando al método getWords() (línea 10). En la línea 11, se realiza un proceso iterativo según las

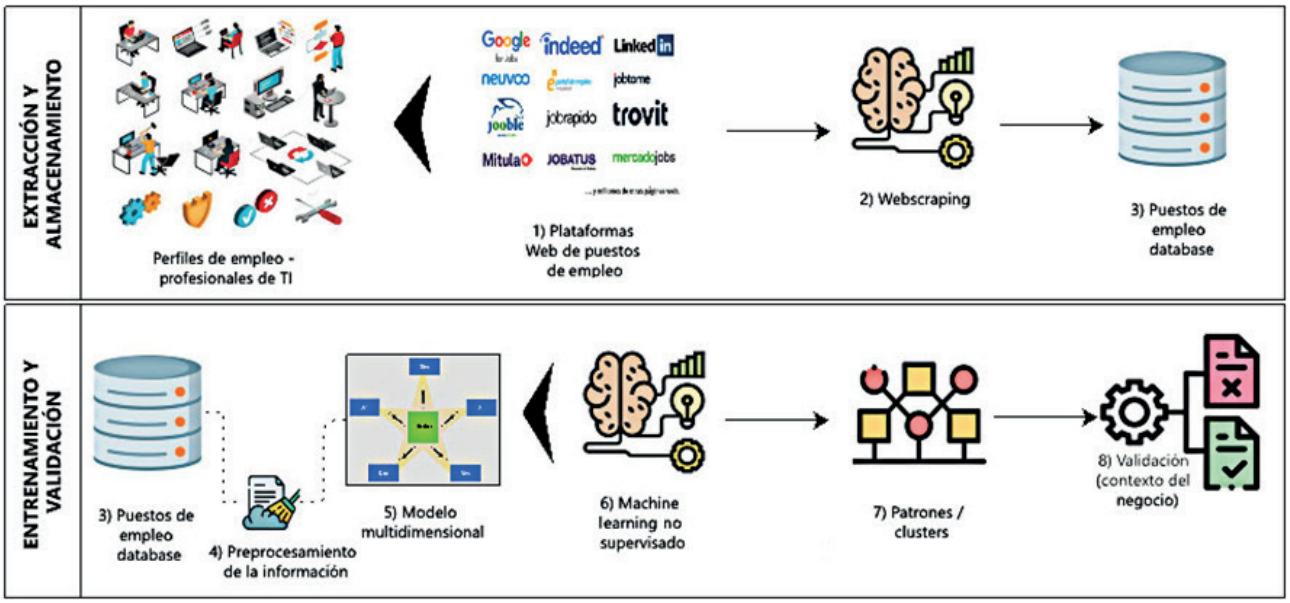

Figura 1. Proceso de *machine learning* no supervisado.

Fuente: Elaboración propia.

Tabla 2. *Extracto del programa para la extracción de puestos de empleo.*

```
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["pa-
gina_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["pagi-
na_inicial"], carga["cant_paginas"], carga["cant_ofertas"],carga["id_carga"])
if __name__ == "__main__":
maledpeti_portal(COMPUTRABAJO)
maledpeti _portal(INDEED)
maledpeti _portal(BUSCOJOBS)
```

Fuente: Elaboración propia.

palabras recuperadas en la línea 10 y por cada palabra realizará las siguientes acciones: inicializar el arreglo carga con los valores establecidos en las constantes WS_PORTAL_LABORAL, WS_PAGINAS, WS_PAGINA_INICIAL, WS_OFERTAS, WS_AREA, las cuales contiene información relativa al portal de trabajo a procesar, el número de páginas a leer, la página iniciar a leer, la cantidad de ofertas o puestos de empleo a leer, así como el área de búsqueda respectivamente, en vista de que un portal de empleo mantiene información de puestos de empleo de distintas áreas disciplinarias. En la línea 20, se setea los parámetros de búsqueda; en la línea 21, se llama al método registrar_webscraping(), que se encuentra definido en la clase *controller*, cuya consecuencia es el poblado de los esquemas de la base de datos: tablas *webscraping* y oferta. Para registrar el detalle del puesto de empleo, se realizan las acciones indicadas en la línea 23, 24 y 26, en función del tipo de portal de trabajo, considerando que los portales describen una estructura html distinta; este procedimiento se repite por cada palabra clave a buscar y considerada en el arreglo palabras (línea 10). En la línea 28, se puede apreciar el método principal *main*, desde el cual se invoca la función maledpeti_portal() pasando como parámetro el portal a procesar con lo cual se pobla la tabla oferta_detalle.

El segundo subproceso concibe el entrenamiento y validación del modelo, este se subdivide en seis actividades, las cuales se detallarán a continuación:

Pre-procesamiento. Esta actividad consiste en aplicar técnicas de minería de texto para la segmentación del detalle del puesto de empleo en subunidades que fueron almacenadas en tuplas individuales en un esquema de la base de datos para su posterior clasificación según el tipo de dimensión que le concierne. El procedimiento seguido en esta etapa es el señalado por Swamynathan (2017, pp. 70, 256): eliminar el ruido, así como los valores atípicos u *outliers* para garantizar un análisis eficiente. Este consistió en limpiar el texto mediante la conversión de los valores de tipo texto a mayúsculas, eliminación de tildes, caracteres extraños, espacios en blanco, textos o palabras sin significados y la tokenización del puesto de empleo. Este último procedimiento consistió en segmentar el detalle del puesto de empleo en componentes significativos simples a los que se asignó el término «dimensión».

En la Tabla 3, se expone un extracto del código en Python que se utilizó para remover el ruido. Por ejemplo, en la línea de código 17, se invoca la función remove_tags_html() para remover las etiquetas

Tabla 3. *Código Python para remoción de ruido.*

```
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]
```

Fuente: Elaboración propia.

html que se encuentren en la cadena de texto; en la línea de código 18, se llama a la función incomplete_tags_html() para remover etiquetas html incompletas que se encuentren en el texto del detalle del puesto de empleo; y, en las líneas 19 y 20, se citan las funciones remove_non_ascii() y remove_space() para remover los códigos no ascii y los espacios en blanco o caracteres en blanco de inicio y fin de la cadena o párrafo de texto correspondiente al detalle del puesto de empleo.

Diseño del modelo multidimensional. Esta actividad se formuló con el propósito de organizar los elementos de información que comprende un puesto de empleo. Se utilizó el esquema copo de nieve característico de un *data mart* de alta granularidad, de acuerdo a lo establecido por Kimball y Ross (2002), el cual está compuesto por una tabla de hechos «puestos de trabajo» y catorce dimensiones principales como son categoría, perfil, portal web, empleador, ciudad, salario, periodo, función, conocimientos, competencias, habilidades, certificaciones, beneficios y formación, como se puede apreciar en la Figura 2.

Modelar la información mediante un esquema multidimensional permitió apreciar la data desde diferentes perspectivas, como se observa en la Figura 3, que muestra el conjunto de *dashboards* del proyecto elaborado con la herramienta para inteligencia de negocios Power BI versión 2.99.862.0. La información concierne a los puestos de empleo de profesionales de TI, la cual ha sido organizada por periodo, página web, empresa, funciones, competencias, habilidades, beneficios y salario; además, es posible incorporar otras dimensiones. Se observan cuatro perspectivas fundamentales, la primera expone a razón de porcentaje los puestos de empleo por página web y se observa que el mayor porcentaje del periodo 2021 corresponde a la página de Google Jobs Search con un 40.18%, mientras que las otras páginas exhiben un porcentaje menor para el mismo periodo.

Asimismo, se aprecia mediante un gráfico circular las habilidades transversales, blandas, sociales y organizativas que suelen requerir los empleadores; se evidencia que la comunicación efectiva, el trabajo en equipo, la capacidad de análisis y la proactividad e iniciativa son las habilidades más demandadas en los perfiles de empleo de profesionales de TI. Por otro lado, se presenta los principales perfiles de TI requeridos por los empleadores como bancos, financieras, mineras, empresas de servicios, retails; en cuanto al sector público, se tiene ministerios, municipalidades, entre otras instituciones gubernamentales. El cuarto *dashboard* expone la demanda de los perfiles por periodo, en donde destacan los perfiles «developer», «soporte técnico», «analista programador», «analista de calidad», «fullstack developer» entre los más demandados en el periodo 2021.

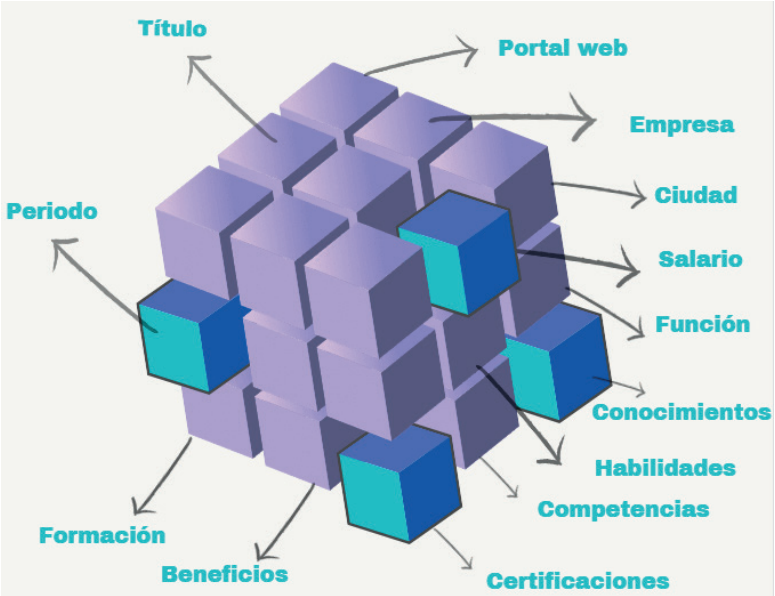

Figura 2. Modelo multidimensional.

Fuente: Elaboración propia.

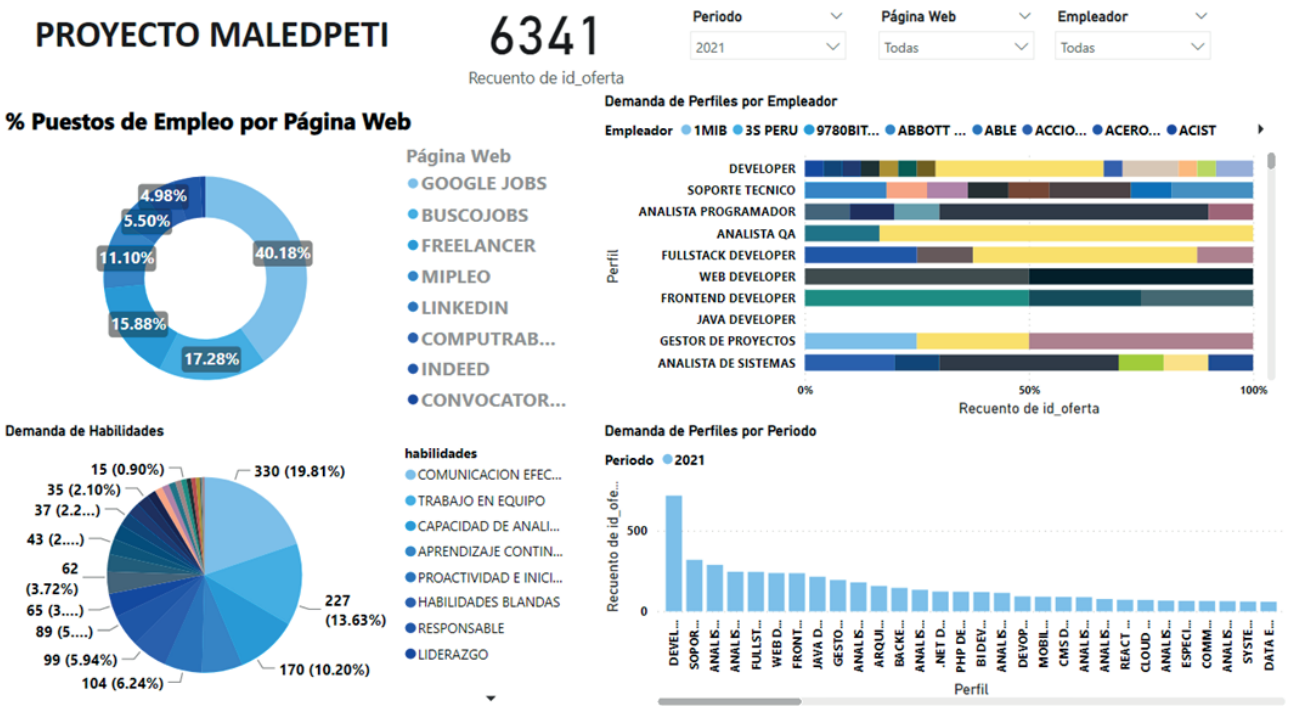

Figura 3. Dashboards del proyecto MALEDPETI.

Fuente: Elaboración propia.

Machine learning no supervisado. Esta etapa comprende la aplicación de la técnica *clustering*. Se trata de una técnica no supervisada, basada en la determinación de conglomerados generados por similitud de puntos respecto a su posición geométrica en el espacio vectorial; es una técnica descriptiva y de clasificación, no está sujeta a ningún modelo formal, no asume la existencia de variables dependientes, ni independientes, no requiere *datasets* entrenados para su análisis; y los modelos se crean automáticamente partiendo del reconocimiento de los datos (Perez, 2014; Swamynathan, 2017; Deshpande, 2018). En Sierra (2006), se define la lógica del algoritmo *k-means* como se indica en la Figura 4.

Complementariamente a las propuestas de Sierra (2006) y Swamynathan (2017), en la literatura se cuenta con diversas variantes del algoritmo *k-means*, estas versiones se enfocan en conseguir la buena calidad de los resultados del *clustering* y en exceptuar la dependencia de establecer el número de *clusters* a obtener como resultado del proceso; entre los incidentes presentados en la aplicación del algoritmo se puede mencionar la no convergencia en los resultados, esto puede deberse al número de *clusters k* elegido o a la ausencia de estructura de *clusters* en los datos según refiere Sierra (2006, p. 290). En la investigación se utilizó la propuesta del algoritmo *k-means*++ de Arthur y Vassilvitskii (2007), esta variante contempla una forma de inicialización basada en la determinación de centroides aleatorios con probabilidades muy específicas que logra mejoras sustanciales en términos de precisión y velocidad respecto al tradicional método *k-means* de Lloyd's. El algoritmo *kmeans*++ requiere que se precise como parámetro de entrada el valor de *k* conglomerados que se desea obtener, por lo que se recurrió a la estrategia denominada método del codo de acuerdo con el trabajo de Nerurkar et al. (2018) para estimar el número ideal de conglomerados *k* en función del punto de inflexión en la gráfica; la función objetivo del algoritmo es minimizar en los conglomerados *k* la suma de las distancias al cuadrado cuya fórmula se puede apreciar en la expresión 1:

$$\text{SSE} = \sum_{k=1}^{k} \sum_{(x_i \in k)} \|x_i - c_k\|^2, \textit{donde } c_k \textit{ es centroide del cluster}. \quad (1)$$

Fuente: Tomado de *Empirical Analysis of Data Clustering Algorithms* (p. 2), por Nerurkar et al. (2018), Procedia Computer Science 125(1).

i. De entre los n casos elegir k que llamaremos semillas y denotaremos c_j , $j = 1, \cdots k$, Cada semilla c_j representará al cluster $C_j (j = 1, \cdots k)$.

ii. Asignar el caso i al cluster C_j cuando $d(x_i, c_j) = \text{mín}_{l=1,\ldots,k,}\, d(x_i, c_l)$. Es decir, cada caso se asigna al cluster que representa la semilla que tiene más cerca.

iii. Los pasos 1 y 2 nos dan una partición inicial de los casos.

iv. Calcular la mejora que se produciría en el criterio elegido (minimizar $tr(W)$, minimizar $det(W)$, etc…) al asignar un caso a otro cluster en el que no está actualmente.

v. Hacer el cambio que mayor mejora produce en el criterio.

vi. Repetir los pasos 3 y 4 hasta que ningún cambio haga mejorar el criterio elegido.

Figura 4. Algoritmo *k-means*.

Nota. La figura expresa la secuencia lógica del algoritmo de tipo *clustering k-means*.
Fuente: Tomado de *Aprendizaje Automático conceptos básicos y avanzados, Aspectos prácticos utilizando el software WEKA* (p. 290), por B. Sierra, 2006, Pearson Prentice Hall.

Se usó la técnica *k-means* con el *software* Weka versión 3.8.5 y se determinó un modelo basado en *clusters*; el algoritmo *k-means* que implementa Weka es el propuesto por Arthur y Vassilvitskii (2007). Desde la interfaz se dispone de varias funciones para el cálculo de las distancias: Chebyshev, Euclidean, Filtered, Manhattan y Minkowski. Otro aspecto relevante es el método de inicialización: Random, k-means++, Canopy y Farthest first, estos parámetros son requeridos por el algoritmo, así como el número de *clusters* a determinar y el número máximo de iteraciones.

En la línea 1 de la Tabla 4, se muestran los parámetros de ejecución del algoritmo en Weka con un *dataset* de catorce dimensiones, 15 clusters, *k-means*++ como método de inicialización y *euclidean* como función distancia. El resultado retornó las siguientes métricas: 7145 instancias evaluadas, 6 iteraciones realizadas y una suma de errores al cuadrado de 34 696.82. Asimismo, en la Tabla 5 se puede apreciar que el *cluster* #5 sería el más significativo al agrupar el mayor número de instancias; el *cluster* expresa que el 17% de los puestos de empleo de profesionales de TI corresponden al año 2020 y al perfil «fullstack developer», publicitado en «Google Jobs Search», para desempeñarse en una empresa de la ciudad de «Lima-Perú», «no especifica» el salario, la función principal que deberá realizar el postulante al empleo es «desarrollo e implementación de proyectos de software», se requiere que el interesado en la plaza evidencie «conocimiento de base de datos», su competencia principal debe ser «desarrollo en lenguaje de programación Java», la «comunicación efectiva» es una de las habilidades blandas importantes para el empleo, el postulante debe contar con «certificación en Scrum», la empresa ofrece «estabilidad laboral», y como formación para el puesto se requiere «egresados o bachilleres de Ingeniería de Sistemas, Informática o afines».

Considerando una reducción de la dimensionalidad del *dataset* a seis, las métricas resultantes se muestran en la línea 2 de la Tabla 4. Estas indican que el algoritmo realizó 4 iteraciones con una suma de error al cuadrado de 16 027.00, lo cual es mucho más bajo respecto al *dataset* de catorce dimensiones. La Tabla 6 expone el *cluster* #3 con 2318 instancias (33%) como el más representativo y señala que el perfil «developer» es el más solicitado por los empleadores, el rol principal que desempeñarían los postulantes a plazas con este perfil sería «desarrollo e implementación de proyectos de software», se requiere contar con «conocimiento de base de datos», se debe evidenciar «competencia

Tabla 4. *Resultados de Clustering K-means con Weka.*

N.°	Técnica	Instancias	N.° dim	N.° Clusters	Método Inic.	N.° Iter.	SE² (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

Fuente: Elaboración propia.

Tabla 5. *Cluster #5 dataset 14 dimensiones.*

Cluster #5 Instancias: 1239 (17%)
Categoría: DEVELOPER
Perfil: FULLSTACK DEVELOPER
Página_web: GOOGLE JOBS SEARCH
Empresa: NO DETALLADO
Lugar: Lima
Salario: NO ESPECIFICADO
Periodo: 2020
Funciones: DESARROLLO E IMPLEMENTACIÓN DE PROYECTOS DE SOFTWARE
Conocimiento: CONOCIMIENTO DE BASE DE DATOS
Competencias: JAVA
Habilidades: COMUNICACIÓN EFECTIVA
Certificaciones: CERTIFICADO EN SCRUM
Beneficio: ESTABILIDAD LABORAL
Formación: EGRESADO O BACHILLER DE INGENIERÍA DE SISTEMAS, INFORMÁTICA O AFINES

Fuente: Elaboración propia.

Tabla 6. *Cluster #3 dataset 6 dimensiones.*

Cluster #3 Instancias: 2318 (33%)
Perfil: DEVELOPER
Funciones: DESARROLLO E IMPLEMENTACIÓN DE PROYECTOS DE SOFTWARE
Conocimiento: CONOCIMIENTO DE BASE DE DATOS
Competencias: JAVA
Habilidades: COMUNICACIÓN EFECTIVA
Beneficio: ESTABILIDAD LABORAL

Fuente: Elaboración propia.

en el lenguaje de programación Java», una de las habilidades blandas más solicitadas es la «comunicación efectiva» y el mayor beneficio ofrecido por parte del empleador es la «estabilidad laboral».

RESULTADOS Y DISCUSIÓN

El diseño de un proceso de ML no supervisado permitió establecer con claridad las etapas a cubrir en la determinación de la demanda social de puestos de empleo de profesionales de TI. El proceso propuesto contempló dos subprocesos: el primero comprende la extracción y almacenamiento de la información concerniente a los perfiles de empleo de profesionales de TI, obtenidas desde las principales plataformas web con el uso de técnicas de *webscraping*, y el segundo subproceso concibió actividades de *clustering* y validación del modelo.

Una vez concluida la fase experimental de la técnica de ML no supervisada *k-means* con el *software* Weka, en la Tabla 4 se mostraron las métricas obtenidas. Considerando el *dataset* de catorce dimensiones, se obtuvo la métrica «suma de errores al cuadrado» más alta respecto al de seis dimensiones; también se aprecia el impacto en el número

de iteraciones requerido para conformar los quince conglomerados o *clusters*, el cual se redujo de seis a cuatro iteraciones.

Modelar la información de los puestos de empleo utilizando un modelo multidimensional permitió obtener diversas perspectivas del comportamiento de la información, una de esas perspectivas corresponde a las funciones o roles requeridos por los perfiles de empleo, parte de los cuales se aprecian en la Figura 5. Una misma función es requerida por varios perfiles, cada color de la barra representa a un perfil, por ejemplo: la función «gestión de proyectos» es una de las labores que desempeñarían los perfiles «web developer», «system administrator», «gestor de proyectos», «developer», «CMS developer», «arquitecto de software», «analista de seguridad informática», «analista QA», «analista Cloud» y «administrador de redes y comunicaciones». Análogamente, se pueden analizar las demás funciones señaladas en la Figura 5.

Otra variable importante en una oferta de empleo son los beneficios ofrecidos por los empleadores, el mayor porcentaje de lo ofrecido corresponde a los derechos de ley según el tipo de la modalidad de la plaza. De los resultados se pudo apreciar que el que más ofrecen es la «estabilidad laboral», seguido por «línea de carrera», «beneficios corporativos», «seguro de salud» y «capacitaciones constantes», mientras que muy pocos empleadores ofrecen «utilidades».

Se ha podido determinar la alta variabilidad de los salarios ofrecidos por las empresas del sector privado y, en menor medida, por las organizaciones públicas. Un gran porcentaje de empresas no precisa el salario en la oferta de empleo, estas estilan señalar «no especificado», «acorde al mercado laboral», «acorde a responsabilidad», «acorde a experiencia y conocimientos», «acorde al proyecto» y «sujeto a evaluación». Otros empleadores sí consignaron los salarios ofrecidos, una muestra de ello son los perfiles «developer» y afines, para los cuales los salarios ofertados oscilan entre <3000 ; 6000> soles.

Por otro lado, las competencias que debe evidenciar el postulante a un puesto de empleo de TI pueden no ser exclusivas de un perfil, una muestra de ello es la competencia técnica «diseñar la arquitectura del software», esta es más demandante en la categoría «developer», pero exigida también en perfiles de las categorías «arquitecto», «gestor»

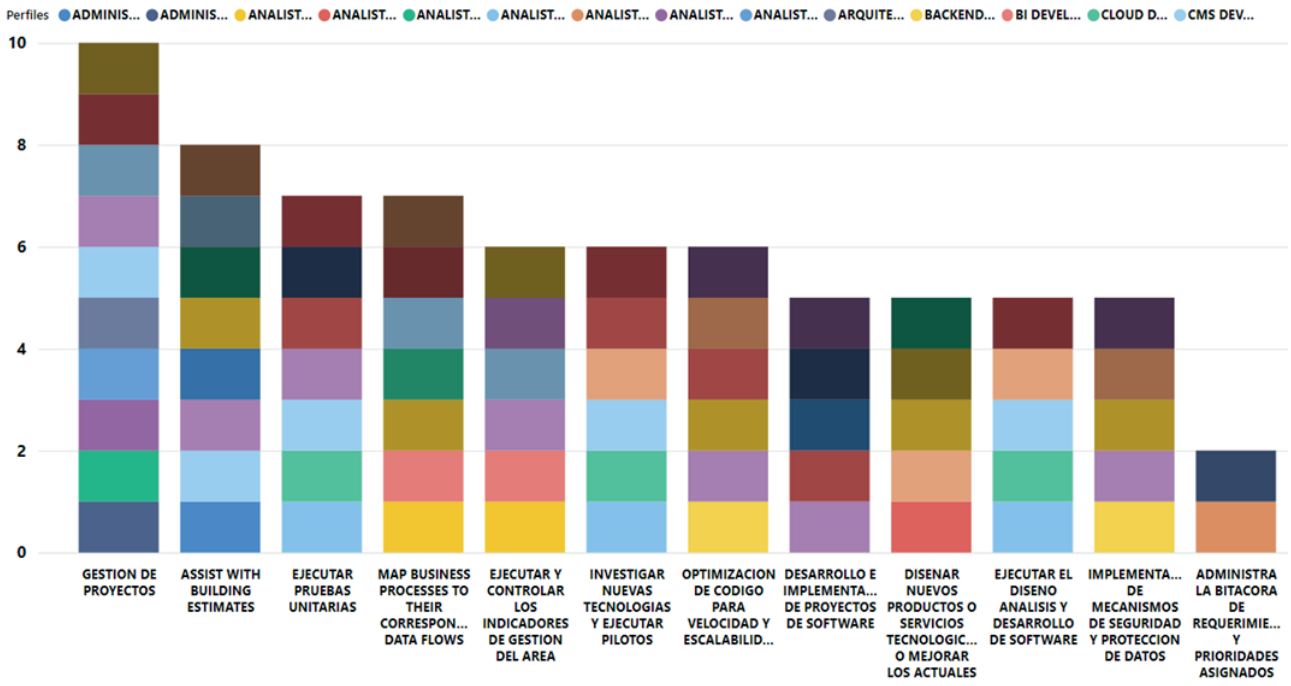

Figura 5. Funciones transversales a perfiles.

Fuente: Elaboración propia.

y con menor énfasis en la categoría «analista de sistemas». Bajo la misma lógica se pueden analizar las demás competencias.

Asimismo, el puesto de empleo del perfil «data engineer» requiere evidenciar cumplimiento de las competencias técnicas como AWS, base de datos, *big data*, *cloud computing*, *data science*, *data warehouse*, desarrollo de *software*, *development platform*, ETL y gestión de procesos. De manera similar, se pueden analizar las competencias exigidas en cada cada perfil de empleo de manera sistematizada.

Las habilidades blandas son otros elementos de información de alto valor en un puesto de empleo, entre las habilidades transversales más demandadas se tiene a la «comunicación efectiva», el «trabajo en equipo» y la «capacidad de análisis»; la «resiliencia» es una habilidad que empieza a exigirse dada la coyuntura actual de pandemia.

CONCLUSIONES

1. La presente investigación plantea un proceso de ML no supervisado para determinar la demanda social de puestos de empleo de profesionales de TI. El proceso propuesto contempla dos subprocesos: el primero comprende la extracción y almacenamiento de la información concerniente a los perfiles de empleo de profesionales de TI, obtenida desde las principales plataformas web con el uso de técnicas de *webscraping* y el segundo subproceso concibe actividades de *clustering* y validación del modelo.

2. Se diseñó un modelo multidimensional compuesto por una tabla de hechos «puestos de empleo» y catorce dimensiones principales como son categoría, perfil, página_web, empleador, ciudad, salario, periodo, función, conocimientos, competencias, habilidades, certificaciones y beneficios, lo que permitió, con el apoyo de una herramienta de inteligencia de negocios, determinar mediante diferentes perspectivas el comportamiento de la demanda social en función de las dimensiones identificadas para tal fin.

3. Se aplicó ML no supervisado mediante el uso de la técnica de *clustering k-means* para determinar del comportamiento de la demanda social, a partir de los conglomerados generados por similitud de puntos respecto a su posición geométrica en el espacio vectorial, con lo que se crearon modelos automáticamente a partir del reconocimiento de los datos, cuyos parámetros utilizados en la ejecución del algoritmo, el número de iteraciones determinados y la métrica intra-cluster se aprecian en la Tabla 4.

4. Las características de los patrones identificados, considerando 14 y 6 dimensiones respectivamente, se muestran en las Tablas 5 y 6, y su significancia se detalla en la sección de resultados y discusión.

5. La propuesta de la presente investigación puede ser considerada en tres escenarios: i) para disponer de una demanda social sistematizada, de interés de actores que toman decisiones en la gestión de programas académicos con fines de licenciamiento y acreditación, ii) para la actualización del Catálogo Nacional de Perfiles Ocupacionales, de acuerdo a la demanda social de las profesiones de TI extendible a otras áreas de conocimiento, y iii) puede ser considerada para un poblamiento sistematizado del repositorio de cualificaciones del Perú.

REFERENCIAS

[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