





Integrando tecnologías asociadas a data science en procesos de orientación vocacional y profesional

Integrating technologies associated with data science in vocational and vocational guidance processes

Yngrid Josefina Melo Quintana¹, Wilson Geovanni Simbaña León ², Andrés Eloy Castillo Rojas ³, Edgar Andrés Bravo Donoso ⁴.

¹ Instituto Tecnológico Universitario Ister, yngrid.melo@ister.edu.ec

² Instituto Tecnológico Universitario Ister wilson.simbana@ister.edu.ec

³ Instituto Tecnológico Universitario Ister andrescastillor@gmail.com

⁴ Instituto Tecnológico Universitario Ister edgar.bravo@ister.edu.ec

Autor para correspondencia: yngrid.melo@ister.edu.ec

Fecha de recepción: 2020.08.14

Fecha de aceptación: 2020.10.16

Fecha de publicación: 2021.01.12

RESUMEN

La Ley Orgánica de Educación Intercultural Bilingüe (LOEI), refiere una flexibilización de los tipos de bachillerato. Según Salgado [1], esperan que se desarrolle un análisis técnico donde haya un papel más fuerte del orientador vocacional. Esto conduce a que la orientación vocacional adquiere mayor importancia. Actualmente se están desarrollando con mucha rapidez y mucha fuerza las tecnologías ligadas a la Ciencia de Datos como la inteligencia artificial (IA) el Big Data y minería de datos (Data mining) para apoyar procesos de tomas de decisiones donde están involucrados grandes cantidades de datos. La orientación vocacional según [2], es un problema de toma de decisiones donde está presente en alto grado la incertidumbre y donde los modelos clásicos de decisión no resultan del todo aplicables para dar solución efectiva.

Se pretende integrar estas herramientas asociadas a la ciencia de datos que ya están aplicándose en campos muy diversos en el proceso de orientación vocacional y profesional. Para ello es necesario contar con datos y por consiguiente el objetivo es concentrarse en su extracción y recolección, así surgen preguntas como

¿qué?, ¿dónde? y ¿con qué? buscar. Además, se describe el papel de estas tecnologías respecto al tema planteado e investigaciones existentes hasta el momento con respecto al mismo. La metodología usada es de tipo documental descriptiva. Se presenta a la par, la propuesta de un proyecto de construcción de una herramienta tecnológica que contribuya a la búsqueda, recolección y clasificación de datos que pueda ser usada en procesos de orientación vocacional y profesional.

Palabras clave: Orientación Vocacional, Big Data, Data Science, Inteligencia Artificial, Minería de Datos.

ABSTRACT

The Organic Law on Bilingual Intercultural Education (LOEI) refers to a relaxation of the types of baccalaureate. According to Salgado [1], they expect a technical analysis to be developed where there is a stronger role for the vocational counsellor. This leads to a more important vocational orientation. Currently, technologies related to Data Science such as Artificial Intelligence (AI) and Big Data Mining (Data Mining) are being developed very quickly and strongly to support decision-making processes where large amounts of data are involved. Vocational guidance according to [2] is a decision-making problem where there is a high degree of uncertainty and where classical decision-making models are not fully applicable to provide an effective solution.

The aim is to integrate these tools associated with data science that are already being applied in very diverse fields into the vocational and vocational guidance process. For this it is necessary to have data and therefore the objective is to concentrate on its extraction and collection, so questions arise as

what? where? and with what? look. In addition, it describes the role of these technologies with respect to the topic raised and research existing so far with respect to it. The methodology used is descriptive documentary. At the same time, the proposal of a project to build a technological tool that contributes to the search, collection and classification of data that can be used in vocational and professional orientation processes is presented.

Key words: Vocational Guidance, Big Data, Data Science, Official Intelligence, Data Mining.

INTRODUCCIÓN

La orientación vocacional y profesional es un proceso que está especialmente ligado a los estudiantes de los últimos años de bachillerato que una vez egresados del nivel de educación secundaria aspiran a proseguir sus estudios de tercer nivel. Este proceso se realiza a través de una serie de pruebas que buscan indagar las aptitudes más desarrolladas en el estudiante con relación a una carrera o profesión. La exploración de sus aficiones y su relación con las diversas áreas del conocimiento permitirán definir un perfil profesional específico o por lo menos muy cercano al adecuado. Desde hace muchísimo tiempo hasta hoy en día, se han buscado y experimentado nuevas formas para mejorar los métodos de asesoramiento vocacional y profesional. Muchos esfuerzos se han desarrollado desde el campo de la educación con el apoyo de las Tecnologías de comunicación (TIC). [2]

La orientación profesional es un problema de toma de decisiones acerca de la conducta vocacional dirigida a una persona o grupo de personas. Ahora bien, la toma de decisiones es una actividad cotidiana de los seres humanos, constantemente las personas se encuentran ante

situaciones donde existen varias alternativas y debe elegir la más conveniente o la que mejor se adapte a sus necesidades. La toma de decisión es en determinadas ocasiones un proceso complejo debido a la necesidad de realizar previamente un análisis detallado de las ventajas e inconvenientes asociados a cada alternativa, y se vuelve más complejo para aquellos problemas donde existe un alto grado de incertidumbre y la información necesaria para decidir es vaga e imprecisa.

Otro problema presente en los procesos de toma de decisiones y más aún en orientación vocacional y profesional es el volumen elevado de información que se necesita para obtener un resultado lo más cercano posible al deseado. Esto aunado a la aparición de las llamadas redes sociales y las páginas web, donde constantemente se generan volúmenes considerables de información relacionadas con gustos, preferencias y aptitudes y un sinnúmero de información de los usuarios de estas redes, datos estos de gran relevancia en lo que determinar la vocación se refiere.

Normalmente los procesos de orientación vocacional son realizados por orientadores que utilizan numerosas pruebas psicológicas llamadas pruebas vocacionales que son aplicadas al estudiante, cada prueba contiene una cantidad considerable de preguntas. Esto genera un volumen extenso de preguntas y por consiguiente innumerables resultados que se pueden conseguir a partir de las respuestas obtenidas. Si a esto se le agrega la información que se puede extraer de internet, estamos hablando de una incuantificable cantidad de datos que cuando se quieren manejar para determinado fin se hace más dificultoso el procesamiento de los mismos, aunado a los grandes costos que esto conlleva.

En este sentido, se han desarrollado muchas aplicaciones con tecnología informática para la orientación y asesoramiento vocacional, muchas de ellas enmarcadas en la teoría clásica de la decisión. Estas herramientas son llamadas CACGS (Computer Assisted Career Guidance Systems) y en su mayoría son programas interactivos, que facilitan la integración de materiales, se prestan a la autoaplicación y autoevaluación, y resultan atractivos para el estudiante. Incluyen componentes de asesoramiento y orientación en el proceso de toma de decisiones, junto a información profesional o bases de datos de carácter educativo. Una muestra de estas herramientas las podemos encontrar en [2], [3], [4]. También se han diseñado diversidad de modelos formales y no formales [5], que miden la capacidad de

decisión del individuo y evalúan su habilidad en dicho proceso. No obstante, los resultados de estas herramientas y modelos son en sentido más exploratorios, limitados y bastantes imprecisos y lo que es más importante, dejan a un lado una gran cantidad de datos importantes y no toman en cuenta el alto grado de incertidumbre que genera tomar una decisión en un aspecto tan importante y que incide totalmente en el futuro de una persona, como es escoger una carrera o profesión.

Siguiendo con lo que concierne a esta investigación, en la actualidad existe una gran variedad de técnicas basadas en herramientas inteligentes, que permiten mejorar los procesos de toma de decisiones, como las técnicas basadas en Inteligencia artificial (IA), la computación blanda (Soft computing), el Big Data, la minería de datos, además de Python como lenguaje de programación que está siendo muy utilizado en procesos que tienen que ver con la ciencia de datos para resolver problemas del manejo y análisis de grandes cantidades de datos para la toma de decisiones. Estas herramientas han provocado en los últimos años un importante número de investigaciones que se encaminan a resolver este tipo de problemas y a mejorar los resultados de dicho proceso, basándose en la rapidez para identificar, procesar y extraer la información que realmente es importante con un mínimo de costos. En el entorno educacional se han utilizado para el desarrollo de métodos que han logrado extraer información útil a partir de los datos que se generan y utilizarla para mejorar dicho entorno. El objetivo de esta investigación es buscar de qué manera todas estas herramientas tecnológicas disponibles actualmente pueden ayudar en el manejo de grandes cantidades de datos y toma de decisiones, en procesos de orientación vocacional y profesional. Estas herramientas y técnicas como Big Data, permiten ir más allá de lo que se conoce de los datos, permiten extraer patrones, son en sí mismas herramientas de investigación. Su uso en procesos vocacionales significa su uso sobre técnicas de investigación sobre datos de gran volumen. El primer proceso para beneficiarse de estas herramientas es disponer de datos, y por lo tanto el objetivo de este proyecto es establecer qué buscar, dónde buscar, con qué buscar y donde almacenar esos datos para sus posteriores procesos en Big Data, por ejemplo. Se espera obtener patrones en los datos, que terminen siendo la base para otros procesos que deriven en tomas de decisiones.

1.1.- Herramientas de Data Science y Orientación Vocacional

Uno de los grandes desafíos de la Inteligencia Artificial (IA) es lograr que una máquina, más específicamente una computadora logre una función similar a la que ocurre en el cerebro. Es así como surge la idea de la computación cognitiva que busca acercarse poco a poco al funcionamiento de este órgano. La IA es una rama de las ciencias de la computación enfocada al diseño y construcción de sistemas, capaces de emular tareas realizadas por un experto humano que engloba muchas ciencias y de ellas se realizan miles de estudios e investigaciones, como el desarrollo de sistemas expertos o sistemas basados en conocimiento enfocados en procesos de orientación vocacional y profesional, como ejemplos se tienen los trabajos presentados por [6], [7] y [8].

Con respecto a otras técnicas de la IA se tiene a bien mencionar a la computación blanda (soft computing), concepto que fue introducido a principios de los años 90 por el Dr. Lotfi Asker Zader, catedrático de la Universidad de Columbia quien fue el primero en publicar, a partir del año 1965, trabajos acerca de los conjuntos difusos y la lógica difusa (fuzzy logic). Básicamente el Dr. Zader expresó que el término computación blanda es una mezcla de distintos métodos y técnicas que cooperan entre sí desde sus fundamentos. El objetivo de la computación blanda es aprovechar la tolerancia que conllevan la imprecisión y la incertidumbre, para conseguir soluciones y una buena representación de la realidad a problemas donde la computación tradicional no opera satisfactoriamente. En [9], se tiene acceso al trabajo original publicado por el autor en cuestión.

Los componentes de la computación blanda son principalmente la lógica difusa que se ocupa de la imprecisión y el razonamiento aproximado, la neuro computación que se emplea en el campo del aprendizaje, y el razonamiento probabilístico que se usa en el manejo de la incertidumbre y la propagación de las creencias [9]. La lógica difusa se ha convertido en un campo muy estudiado en diferentes niveles y usado en diferentes aplicaciones del mundo real en las que aparece incertidumbre, también es utilizada en conjunto con la computación con palabras (Computing With Word), donde se puede desarrollar un tipo de computación más cognitiva [10]. Un ejemplo del uso actual de la lógica difusa se puede encontrar en [11].

Ahora bien, en los procesos de orientación vocacional y profesional están presentes todos los aspectos que se han venido mencionando como parte del tipo de problemas a resolver como son: alto grado de incertidumbre, grandes volúmenes de información y elevados costos

asociados al manejo de la calidad de la toma de decisiones. Es por ello que las técnicas y ciencias de la IA mencionadas se pueden aplicar como ayuda o soporte a este proceso tan importante dentro de la etapa de evolución del ser humano. Pero además existen otros métodos y herramientas dentro de la Inteligencia artificial, que actualmente están siendo muy utilizadas en el manejo de grandes cantidades de datos para procesos de tomas de decisión, como es la Ciencia de Datos o Data Science y dentro de esta, la Big Data la cual se han convertido en tendencia mundial, dentro de esta ciencia. [12]

La Ciencia de Datos es según [12] un concepto relativamente nuevo que se usa en forma general para referenciar a la cadena de técnicas utilizadas para el tratamiento y manipulación de información masiva desde dos enfoques; el estadístico y el informático. Con esto surgió también un nuevo perfil profesional, el “Data Scientist”, son las personas dentro de este perfil que saben del negocio, las herramientas computacionales y del análisis e interpretación estadística. Data Science es un campo que abarca limpieza, preparación y análisis de datos. Es también empleada en muchos campos científicos como las matemáticas, las estadísticas y muchas herramientas más que se aplican a un conjunto de datos para extraer conocimiento de los mismos.

La Data Science revela tendencias y genera información que los investigadores pueden utilizar para mejorar los procesos de toma de decisiones, así como para crear productos y servicios más innovadores en este campo. Algunas tareas dentro de la ciencia de datos están: desarrollar estrategias para analizar datos; la preparación de datos para su análisis; la exploración y visualización de datos; la construcción de modelos con datos mediante lenguajes de programación como Python y R; y la implementación de estos modelos en aplicaciones. [13]

Por otro lado, el término Big Data está asociado esencialmente a grandes cantidades de datos, según [14] hoy en día se generan diariamente cerca de 4,5 cuatrillones de bytes de datos, la mayor cantidad en internet a través de páginas web y las RRSS, que pueden ser utilizados para diferentes análisis, pero para ello se necesita hacer una gestión correcta de los mismos. Es allí donde obligatoriamente se tienen que disponer de sistemas capaces de analizar estos datos para extraer información conveniente y utilizable. A raíz de esto es que surge el término analítica de datos, que va desde la invención del business Intelligence (BI) hasta el Big Data.

El objetivo es analizar grandes cantidades de datos tanto estructurados como no estructurados y obtener información para mejorar la toma de decisiones. Pero a pesar de que Big Data se asocia principalmente con cantidades de datos exorbitantes, también abarca tanto volumen como variedad de datos y velocidad de acceso y procesamiento.

Paralelamente a la analítica de datos surge otra técnica para el procesamiento de datos: la minería de datos o data mining, que consiste básicamente en buscar y extraer información de un gran conjunto de datos a través de patrones y modelos. Según [14] la minería de datos se basa en un proceso que intenta descubrir patrones o comportamiento de información en un conjunto masivo de datos y que no requiere una cantidad de datos específica pero sí de datos relevantes. El data mining tiene capacidad para manejar una infinidad de datos, por lo que puede decirse que la minería de datos y el Big Data van de la mano.

La tecnología de Big Data es información con gran volumen, velocidad rápida y mucha variedad que demanda una plataforma innovadora para mejorar los conocimientos para la toma de decisiones, es una forma de resolver todos los problemas no resueltos relacionados con la gestión y el manejo de datos. Con la analítica de datos y el data mining se puede reconocer patrones ocultos y conocer una visión de las personas y comprender mejor sus necesidades.

1.2.- Python y la ciencia de datos

Como se ha expresado en párrafos anteriores, mediante modelos predictivos de minería de datos es posible identificar características y patrones relacionadas con las aptitudes y actitudes de las personas a través de los medios de comunicación y redes sociales que utiliza diariamente. En este sentido, uno de los lenguajes de programación más usados recientemente en desarrollo de estrategias para analizar, preparar, explorar y visualizar datos; así como para construir e implementar modelos en aplicaciones con datos es Python, aunque tiene un gran competidor, el lenguaje de programación R. Según [15] la diferencia entre Python y R es que este último es un lenguaje de dominio específico, orientado específicamente al análisis estadístico y, en consecuencia, su construcción está pensada y dirigida a ese uso. Por otro lado, señalan los autores que Python es un lenguaje de uso más general, a la vez más rápido y seguro que R, y cualquier algoritmo es más fácil de desarrollar.

La experiencia del uso del lenguaje de programación Python en la ciencia de datos, como se manifiesta en [16], abarca proyectos de software de diferente magnitud, que van desde pequeñas pruebas exploratorias o estudios que extraen conocimiento de los datos con unas pocas líneas de código, a proyectos a gran envergadura, abordados por equipos de trabajo multidisciplinarios para lograr este mismo propósito.

Python se puede utilizar para abarcar todas las estrategias de: captura de datos, donde se incluyen la interacción con fuentes de datos de terceros a través de APIs así como librerías específicas que permiten recolectar datos de cualquier web; de igual forma, librerías que proporcionan estructuras de datos y las herramientas de tratamiento de datos como la muy utilizada librería pandas; librerías que implementan algoritmos de aprendizaje automático como scikit-learn; igualmente incluye librerías de visualización de datos genéricas como matplotlib y especializadas para varios tipos de datos como los grafos con networkx o mapas con congeoplotlib. [16]

En recapitulación, la gestión de los datos según [17] se compone de tres aspectos: en primer lugar está la adquisición y almacenamiento de los datos, en segundo lugar la limpieza y depuración de los datos y, por último la preparación para su análisis. Por otra parte, la analítica de datos está referida a la respuesta de preguntas formuladas a partir de técnicas de modelamiento y análisis. Como se puede evidenciar, estos procesos no son distintos del proceso de indagación científica presente en cualquier disciplina, la diferencia reside en las características principales de los datos que se utilizan y los retos que significan el acceso y manipulación de los datos para cualquier fin. Trabajos como los presentados en [18],[19] y [20] entre otros, han demostrado que mediante el análisis de grandes cantidades de datos se pueden obtener respuestas que antes no eran conocidas en el sector educativo, prometiendo nuevas perspectivas que aportan mucha atención en diferentes campos del aprendizaje. Varias de estas técnicas se podrían aplicar en los departamentos de orientación, obteniendo nuevas herramientas para trabajar con los alumnos la parte de orientación vocacional y profesional.

La investigación aquí presentada, tiene como principal objetivo establecer la extracción y recolección de datos para procesos vocacionales y profesionales por los que buscar, donde buscar, conque buscarlos y donde almacenarlos son los elementos a tener en cuenta.

La exploración de literatura sobre el uso de Data Science y sus diferentes técnicas y herramientas en procesos asociados a la orientación vocacional y profesional, guiará hacia la propuesta de un software basado en estas técnicas para recolección y clasificación de datos que puedan ser usados posteriormente para la toma de decisiones con respecto a la escogencia de una carrera profesional en base a la vocación del individuo.

MARCO TEÓRICO

La metodología empleada para la exploración incluye la aplicación de dos estrategias: primeramente, de tipo descriptiva con diseño documental, basada en la revisión de literatura relacionada con el tema. La revisión de literatura se adapta al estudio aquí presentado para conocer el estado de las investigaciones sobre el uso de las técnicas de ciencia de datos en procesos de orientación vocacional. Esto será el punto de partida para la realización de proyectos futuros en este ámbito.

Se aplicó una metodología con un enfoque descriptivo y de tipo documental, para lo cual se investigaron distintas fuentes bibliográficas y distintas bases de datos científicas como Google académico, Base, Dialnet, Scielo, Refseek y Redalix, así como en Scopus, las cuales contienen una colección bastante representativa y muy completa.

Luego de la revisión bibliográfica y de plasmar el resultado de la misma, se expondrán aspectos relacionados a la propuesta de un software utilizando técnicas de Big Data para la extracción, recolección y clasificación de datos que puedan luego ser utilizados en una herramienta tecnológica para obtener resultados en lo que a vocación profesional se refiere. En referencia a esto, en todo software que se desarrolle para poder analizar, clasificar y obtener información a partir de datos, lo fundamental es contar con los datos, y cuidar de la calidad cuando se trabaja con un gran volumen de datos puede facilitar la generación de modelos para el análisis [21]. Es por ello que la propuesta se encamina principalmente a este objetivo, la extracción, clasificación y almacenamiento de los datos, a partir de allí se pueden aplicar técnicas de minería de datos enfocadas en patrones para obtener las respuestas buscadas.

Por consiguiente, la segunda estrategia es relativa a la propuesta de una herramienta tecnológica para la extracción, recolección, clasificación y almacenamiento de datos concernientes a estudios de orientación vocacional y profesional. Según [15] el proceso de utilización de Big Data inicia identificando el problema y los datos que lo representan. Se debe suponer qué datos se necesitan, dónde se pueden encontrar y cómo conseguirlos. Una vez que se tiene disponibilidad de los datos, se seleccionan aquellos que sean útiles según los objetivos propuestos y se preparan colocándolos en un formato adecuado. Respecto a que datos se necesitan es dar respuesta a la pregunta ¿Qué buscar? en este particular serían los elementos o datos que se pretenden obtener, extraer y coleccionar, para ello se aplicará el estudio mapeo sistemático, para determinar conceptos y para hallar la literatura que contiene el estado del arte que en principio son fuentes de extracción de datos. Seguidamente se presenta la interrogante ¿dónde buscar?, en este sentido los estudios de mapeo sistemáticos deberían ofrecer un cumulo de documentos relevantes, además la web es una fuente importante de datos. La primera aproximación de búsqueda serán fuentes PDF y WEB del propio internet. Por último ¿Con qué buscar?, para ello, la primera herramienta a usar es Python. Para la búsqueda y extracción se utilizara herramientas de scraping disponibles en el mercado y crawl basado en Python, el almacenamiento se prevé en MongoDB, una base de datos distribuida basada en documentos que almacena datos en formato JSON y contiene un potente lenguaje de consulta [22].

MATERIALES Y MÉTODOS

Se ejecutó en primer lugar, un proceso de revisión de trabajos y artículos científicos que permitió obtener información referente a las herramientas de ciencia de datos aplicadas en entornos educativos. Se logró identificar en algunas de las principales fuentes bibliográficas conocidas, que la mayoría de los estudios y proyectos de uso de técnicas como Big Data y minería de datos, así como de IA, están orientados al problema académico de la deserción escolar y por ende asociadas también a facilitar el estudio del rendimiento académico.

Se encontraron sólo 2 trabajos en referencia al uso de estas herramientas en procesos de orientación vocacional y profesional. Esto significa que es un campo de estudios poco investigado y el cual se puede explotar mucho, de allí la importancia del presente artículo.

Se presentan a continuación estos trabajos de tesis, uno de maestría y otro de pregrado que hacen mención al uso de estas técnicas en procesos de orientación vocacional y profesional, y que se hace necesario destacar en este artículo, tales trabajos fueron:

Estudio del desempeño académico de estudiantes colombianos en las pruebas saber 11 y saber pro para la elección vocacional y permanencia universitaria [23]. Este trabajo de investigación elabora un análisis sobre el desempeño de los estudiantes en las pruebas Saber 11 y Saber Pro y su relación con la elección vocacional, mediante técnicas de Ciencia de Datos. Se emplearon modelos como: 1) correlaciones para identificar las asociaciones entre las distintas variables de las Pruebas Saber 11 y Saber Pro, con el rendimiento académico; 2) Clustering con K-means para realizar agrupaciones que facilitan el perfilado de los estudiantes; 3) Reglas de asociación con “Apriori” para fortalecer las interpretaciones de las relaciones de asociación y consecuencias entre los datos. Se tomaron en cuenta múltiples datos que luego de ser depurados, permitieron extraer información a partir de las observaciones de 132,067 estudiantes sobre los cuales se exploraron distintas características de tipo socioeconómico, académicas, vocacionales y hasta de condiciones especiales como limitaciones físicas o emocionales o pertenencia a grupos étnicos o minorías. Por último, la autora manifiesta que el conocimiento construido mediante las técnicas de ciencia de datos servirá de base para la creación de una prueba de orientación vocacional adecuado al país y a su sistema de educación superior.

RESULTADOS Y DISCUSIÓN

Desarrollo de una herramienta de asistencia para el análisis de pruebas psicométricas de una población grande utilizando técnicas de Big Data. [24] Se genera una herramienta que asiste a los psicómetras en el proceso de valorar una población muy grande, requiriendo de las modernas técnicas de Big Data. En este trabajo se desarrolló una metodología que clasificó los datos recogidos de test psicométricos estandarizados, con el propósito de ser analizados en el contexto del Big Data, para que sirvieran de soporte a los psicólogos de diferentes áreas, para ayudar a tomar decisiones, hacer estudios o investigaciones en grandes poblaciones. Las técnicas utilizadas fueron principalmente creación de algoritmos y utilización de librerías con el lenguaje Python con métodos como: Clustering, K-Means y THE CURE.

Como se evidencia no hay mucho material investigativo aún en esta temática de combinar las técnicas de ciencia de datos con los procesos de orientación vocacional y profesional. Es así como en el marco de esta investigación, se puede inferir que se emplean múltiples algoritmos de minería de datos para estimar aspectos tales como el rendimiento académico y de deserción escolar en todos los niveles educativos en diversos países, pero muy pocos dirigidos hacia el estudio de la orientación vocacional y profesional. En este sentido, la motivación del trabajo de investigación, es desarrollar modelos y herramientas de software basados en Big data y minería de datos que permitan pronosticar la vocación de una persona en el aspecto profesional.

Dentro de una arquitectura Big Data, una de las partes más importantes es donde se llevan a cabo los procesos ETL (extracción, transformación y carga) que es la parte en la que se enfoca la herramienta propuesta. La extracción de datos según Hernández citado en [25], es el proceso para la recuperación de datos de diversas fuentes. Con regularidad, se extraen datos para procesarlos más, migrar los datos a un repositorio de datos o para analizarlos más a fondo. Es también común transformar los datos como parte del proceso.

Como se evidencia, el estudio de la orientación vocacional y profesional utilizando herramientas de ciencia de datos está aún iniciándose, cualquier aporte en este sentido es bienvenido. Por ejemplo como señala [26] el uso del Big Data puede mapear caminos más claros para desarrollar las habilidades e intereses más fuertes de cada estudiante que le permitan repensar su vocación, con esto se podría reducir sustancialmente la deserción y cambio de carrera durante los estudios.

Se propone así una herramienta tecnológica con técnicas de Big Data para extraer, recolectar, clasificar y almacenar datos que tengan incidencia en estudios asociados al proceso de orientación vocacional. Dicha herramienta será dividida en 4 módulos: Búsqueda, extracción, recolección y almacenamiento, pero para ello hay que primero tener claro 4 aspectos fundamentales:

- 1.- ¿qué datos se necesitan? El problema de escoger los datos a extraer en una tarea fundamental, aquí se trata de adaptar la herramienta Big Data al tema que interesa y no tanto por las grandes cantidades de datos que se pueden estar manejando sino por la cantidad de fuentes, con ausencia de formato común y la capacidad de analizar y encontrar patrones

ocultos en la información. En el entorno vocacional y profesional ahora se cuenta con nuevas carreras y profesiones, portafolios personales que demuestran competencias particulares. La forma tradicional en que psicólogos y orientadores han evaluado los pensamientos, los sentimientos y los rasgos de personalidad ha sido mediante cuestionarios, pero como expresan [27], existen nuevas alternativas como el análisis de contenido conducido por los datos y el enfoque del vocabulario abierto utilizando grandes volúmenes de información disponibles en las redes sociales sobre aspectos familiares, socioeconómicos y académicos de los estudiantes, a manera de conocer sus antecedentes así como lo referente a su autoestima, de asertividad, de estilos de aprendizaje y de habilidades de estudio.

2.- ¿dónde se pueden encontrar? Respecto a donde buscar, la gran mayoría de la información recopilada sobre las personas en Internet se deriva de las búsquedas y de las interacciones en redes sociales. Diariamente las personas generan una cantidad masiva de datos en diferentes formatos: audio, imágenes, texto y videos que tratados correctamente pueden ser utilizados para resolver una infinidad de problemas, entre ellos el mencionado en este artículo. Esa información se puede encontrar en servidores de Google, Facebook, entre otras empresas, en este sentido como señala [26] una de las discusiones más importantes es la necesidad de disponer de esos datos para que científicos y sociólogos puedan mapear la conducta de los usuarios, de manera que puedan ofrecer soluciones concretas a problemas humanos como su vocación profesional.

3.- ¿qué se necesita para buscar los datos? para búsqueda en la web de manera automática se dispone de herramientas para realizar web scraping, como bibliotecas de programación, aplicaciones de escritorio, extensiones de navegadores y aplicaciones web, no es objetivo de este artículo analizar cada uno de dichas herramientas, pero por conocimiento y experiencia de los autores, para el software propuesto se propone el uso de Scrapy, un framework de programación para Python, que cuenta con múltiples funciones para realizar web scraping. Con esta herramienta se puede realizar crawling para la navegación automática de sitios web por medio de los hipervínculos por lo cual se puede seguir realizando scraping a otras páginas [28]. También se dispone de librerías en Python como Pandas para la manipulación y limpieza de los datos

4.- ¿dónde almacenar? Para el almacenamiento de datos extraídos de técnicas de Big Data se usan las bases de datos NoSQL, que son las recomendadas cuando se generan grandes volúmenes de datos y con diferentes formatos. Para el módulo de almacenamiento de la propuesta se utilizará MongoDB, que es una de las bases de datos documentales más utilizada actualmente, que permite además la manipulación de los datos a través de Apis orientadas a objetos. Para más detalles acerca de MongoDB se puede consultar la web.

CONCLUSIONES

La ciencia de datos es hoy en día una herramienta fundamental para la explotación de datos y la generación de conocimiento. Entre los objetivos que persigue se encuentra la búsqueda de modelos que describan patrones y comportamientos a partir de los datos, con el fin de tomar decisiones o hacer predicciones. Es un área que ha experimentado un enorme crecimiento al extenderse el acceso a grandes volúmenes de datos e incluso su tratamiento en tiempo real. Este artículo concertó las investigaciones de autores sobre conceptos asociados a la ciencia de datos y los aportes más importantes del uso de herramientas asociadas al análisis de datos en procesos de orientación vocacional y profesional.

Mediante la revisión sistemática se logró identificar algunos problemas que en el ámbito educacional se intentan resolver con los procesos de ciencia de datos, entre los cuales se pueden mencionar el rendimiento académico de los estudiantes, la deserción y la calidad de las estrategias educativas, entre otros. Solo se encontraron dos estudios relacionados con el proceso de orientación vocacional y profesional empleando técnicas de minería de datos.

El desarrollo de la investigación contribuye a la revelación de un área de investigación poco explorada y del cual se pueden generar muchos proyectos e investigaciones científicas que permitan dar respuesta a la problemática planteada acerca de los procesos de orientación vocacional y profesional.

La realización de la propuesta de software para la extracción, recolección, clasificación y almacenamiento de datos relevantes para los procesos de orientación vocacional y profesional, y la culminación de forma satisfactoria de este software, será de gran beneficio para proyectos donde se necesiten estos datos para su posterior análisis, evaluación,

visualización y generación de resultados. El objetivo final es que se pueda utilizar la información generada en algún sistema informático donde se puedan tomar decisiones a partir de los resultados obtenidos.

REFERENCIAS

- ¿Qué cambios trae la reforma a la Ley de Educación?, in El Comercio. 2021.
- Santana Gustavo, V.J.A., Hacia un Sistema Virtual de orientación vocacional. Revista Cubana de Educación Superior, 2019.
- Salas Irina, A.C., Orue Goar, La orientación vocacional en la Enseñanza Media Superior vista desde un sistema informático experto. EduSol, 2020. 20: p. 41-56.
- Chávez, C.F., Sistema experto web para medir el perfil vocacional de los postulantes del centro pre universitario a la UNASAM, in Facultad de ciencias. 2020, Universidad nacional Santiago Antunez de Maloyo Mexico.
- Cabrera, M.C., Models of vocational orientation for the choice of university training careers, in Revista Iberoamericana De La educación. 2020.
- Meza Julio, R.L., Rico Ricardo, Estrada Yadira, Rodríguez Nadia, Diseño de un prototipo de sistema experto que permita identificar patrones basado en las aptitudes y habilidades académicas de los alumnos de nivel superior, in Revista Aristas: Investigación Básica y Aplicada. . 2020, Revista Aristas: Investigación Básica y Aplicada. : México.
- Aguirre Andrea Michael , C.G.J., Sistema de recomendación de carreras profesionales aplicando técnicas de learning analytics e inteligencia artificial para los estudiantes de bachillerato de la Unidad Educativa Particular Bilingüe Ecomundo., in Facultad de Ciencias Matemáticas y Físicas. 2020, Universidad de Guayaquil: Guayaquil, Ecuador.
- Rojas Rafael, H.H., Arauco Saúl Ernesto, Rojas John Fredy Neural Network for Vocational Guidance Based on The Applicant's Profile For Admission To A Study Program, in Turkish Journal of Computer and Mathematics Education. 2021: Turquía.
- Zadeh, L.A., Fuzzy logic, neural networks, and soft computing, in Communications of the ACM. 1994. p. 77-84
- Cotaro, S.G., Contribución al estudio de la computación con palabras/percepciones, in Departamento de Inteligencia Artificial. 2007, Universidad Politécnica de Madrid: Madrid.
- Castellanos Jose Ulises, P.C., Gutierrez Luis Carlos Modelo basado en lógica difusa para la construcción de condiciones de alta calidad en el sistema educativo, in CITAS: ciencia, innovación, tecnología, ambiente y sociedad. 2020.
- Hernández Emilcy, D.N., Moreno Julián., Big Data: una exploración de investigaciones, tecnologías y casos de aplicación, in TecnoLógicas. 2017, Instituto Tecnológico Metropolitano: Colombia. p. 15-38.
- Oracle. ¿Qué es la ciencia de datos? 2020 [cited 2021 15]; Available from: <https://www.oracle.com/mx/data-science/what-is-data-science/>.
- Castillo, J.A., Big Data IFCT128PO, I. Editorial, Editor. 2018: Málaga.

- García Jesús, M.J., Berlanga Antonio, Patricio Miguel, Bustamante Álvaro, Padilla Washington, Ciencia de datos. Técnicas analíticas y aprendizaje estadístico, AlfaOmega, Editor. 2018, Publicaciones Altaria, S.L.: Colombia.
- Jiménez Brian , P.C., Balado Pau Andrio, Marco M. Jesús., Aprendiendo a programar. Nuevos retos, nuevas propuestas, in Actas de las Jenu. 2019: Barcelona, España. p. 71-78.
- Rodríguez Patricio, P.N., Mondaca Javier El uso de datos masivos y sus técnicas analíticas para el diseño e implementación de políticas públicas en Latinoamérica y el Caribe Banco Interamericano de Desarrollo 2017.
- Camargo, A.J., Modelo para la predicción de la deserción de estudiantes de pregrado, basado en técnicas de minería de datos. 2020, Universidad de la Costa: Colombia.
- Quiñonez Lenin, C.Y., Rendimiento académico empleando minería de datos, in Revista Espacios. 2020: Perú.
- Beguerí Graciela, M.A., Minería de datos y una aplicación en la educación superior, in XIX Workshop de Investigadores en Ciencias de la Computación (WICC 2017, ITBA, Buenos Aires). 2017 Red de Universidades con Carreras en Informática (RedUNCI).
- Russo, J.P. Mapeo sistemático y evaluación de arquitecturas de software para contextos de big data. 2018.
- MongoDB, I. La base de datos líder para aplicaciones modernas. 2021; Available from: <https://www.mongodb.com/es>.
- Moreno, L.N., Estudio del desempeño académico de estudiantes colombianos en las pruebas saber 11 y saber pro para la elección vocacional y permanencia universitaria, in Facultad de Ingeniería en Tecnologías de Información y Comunicación. 2019, Universidad Pontificia Bolivariana: Colombia.
- Ardila, S., Desarrollo de una herramienta de asistencia para el análisis de pruebas psicométricas de una población grande utilizando técnicas de Big Data., in Programa de Ingeniería de Sistemas Y Computación. 2017, Universidad Tecnológica de Pereira: Colombia.
- Nazareno Maximiliano , P.M.Á., Análisis de herramientas de extracción de comentarios de la red social Twitter y creación de un corpus no relacional para el posterior uso con el Procesamiento de Lenguaje Natural (PLN) y Minería de Opiniones (MO), in Facultad de ciencias matemáticas y físicas. 2021, Universidad de Guayaquil: Guayaquil, Ecuador. p. 143.
- García-Bullé, S. Big data en el futuro de la innovación educativa. 2019; Available from: <https://observatorio.tec.mx/edu-news/bigdata-educacion>.
- Mariñelarena-Dondena L., E., M. ,Castro Solano A., Extracción de conocimiento con técnicas de minería de textos aplicadas a la psicología, in Revista Argentina de Ciencias del Comportamiento. 2017: Argentina. p. 65-76.
- Moo Armando, A.R., Hernández Francisco, Herramientas Web Scraping para la Adquisición de Datos Turísticos, in KIKAME. 2020: Tepic, Nayarit.