www.ecuadorencifras.gob.ec

CUADERNO DE TRABAJO NO. 17
ABRIL 2024

EL MERCADO LABORAL ECUATORIANO EN LA ERA DIGITAL:

transformar los datos de vacantes de empleo en perspectivas estratégicas a través de la analítica avanzada y machine learning

INEC
Buenas cifras, **mejores vidas**

Autoridades:

Roberto Castillo A.
Director Ejecutivo

Jorge García-Guerrero
Subdirector General

Darío Vélez J.
Coordinador General Técnico de Innovación en Métricas y Análisis de la Información

Carmen Granda E.
Directora de Estudios y Análisis de la Información

Autores:

Diego Del Pozo V.
Instituto Nacional de Estadística y Censos, Ecuador

Andrés Villacís M.
Instituto Nacional de Estadística y Censos, Ecuador

Los Cuadernos de Trabajo Temáticos son documentos que presentan análisis de fenómenos sociales, económicos y ambientales con el objetivo de promover la investigación e incentivar el debate.

El mercado laboral ecuatoriano en la era digital: transformar los datos de vacantes de empleo en perspectivas estratégicas a través de la analítica avanzada y machine learning

Diego Del Pozo V.[1]; Andrés Villacís M.[1]

Resumen

Los rápidos cambios en el mercado laboral ocasionados por el avance de la tecnología y la globalización, hacen esencial mejorar el emparejamiento entre trabajadores y empleadores para ocupar un puesto de trabajo. Esto requiere la reducción de la brecha de información entre ambas partes para responder adecuadamente a la creciente demanda laboral actual. Para hacer frente a este reto, se ha desarrollado un innovador software que utiliza técnicas de *web scraping* para extraer datos de los principales portales de empleo de Ecuador, como Computrabajo (2023), Multitrabajos (2023) y Encuentraempleo (2023). El software utiliza técnicas avanzadas de minería de texto que emplean arquitecturas basadas en transformadores, como *all-mpnet-base*-v2 y *Sentence-T5-Large*, para estandarizar y normalizar la información contenida en las ofertas de empleo. Además, el trabajo implementa el algoritmo ensamble, *Extreme Gradient Boosting* (XGBoost) para imputar los datos que faltan en las variables de educación y experiencia. La herramienta resultante es fiable y coherente, reduce las asimetrías de información en el mercado laboral y permite contar con información relevante para la toma de decisiones de política pública.

Palabras clave: vacantes laborales, machine learning, natural language processing, web scraping

Introducción

El mercado laboral ha sufrido transformaciones significativas impulsadas por la globalización, avances tecnológicos y reorganización de los procesos productivos. Estos cambios han originado una evolución en la demanda de competencias y cualificaciones laborales, resultando en la obsolescencia de ciertos empleos y el surgimiento de nuevos roles (Colombo et al., 2018). Esta dinámica ha provocado un desajuste entre la oferta y la demanda de trabajo, afectando especialmente a países en desarrollo, donde se evidencia un impacto negativo en la estabilidad económica y social (Benítez et al., 2016).

La causa principal de este desajuste es la falta de mecanismos efectivos para recopilar y proveer información precisa sobre las características y requerimientos de los puestos de trabajo. Esta deficiencia conduce a una asimetría de información en el mercado laboral (Cárdenas, 2020). No obstante, el progreso en tecnologías de información ha facilitado el desarrollo de herramientas como son el *web crawling* y *web scraping*. Estas herramientas permiten capturar datos de portales de empleo en línea, posibilitando un análisis estadístico detallado de las vacantes laborales y ayudando a reducir la brecha informativa entre empleadores y trabajadores (Cárdenas et al., 2015).

En ese sentido, este trabajo presenta una metodología innovadora para recopilar información sobre la demanda laboral insatisfecha en Ecuador, mediante el desarrollo de un software avanzado que integra técnicas de Big Data y *web scraping*. El software está diseñado para analizar las ofertas de empleo disponibles en las principales bolsas de trabajo en línea de Ecuador, como Computrabajo (2023), Multitrabajos (2023) y Encuentraempleo (2023). La información recolectada se unifica en una base de datos central, que luego es depurada y normalizada para crear distintas variables de interés a partir de los datos obtenidos de los portales de empleo. Dentro de las variables generadas se abarcan diversos aspectos, tales como el nombre de la empresa, el cargo ofrecido, el género y edad del candidato, el tipo de contrato, la jornada laboral, el salario ofrecido, la ubicación geográfica del empleo, los requisitos

[1] Los investigadores pertenecen a la Dirección de Estudios y Análisis de la Información del INEC. Para requerimiento de información del estudio diríjase a: inec@inec.gob.ec.

educativos, la experiencia solicitada, el número de vacantes disponibles, entre otras. Este enfoque permite obtener una visión integral y detallada de la demanda laboral en el país, facilitando la identificación de tendencias y necesidades específicas del mercado laboral ecuatoriano.

Una vez la base de datos ha sido construida, a través de la minería de texto y algoritmos que emplean modelos avanzados de similitud textual, como all-mpnet-base-v2 y Sentence-T5-Large, el sistema estandariza los datos recabados según clasificaciones armonizadas de interés. Esto incluye la División Político Administrativa (DPA) de Ecuador (INEC, 2022) para la ubicación geográfica de los anuncios de empleo, la Clasificación Internacional Industrial Uniforme de todas las Actividades Económicas (CIIU.4.0) (INEC, 2012a) para una detallada segmentación de sectores empresariales, y la Clasificación Internacional Uniforme de Ocupaciones (CIUO.08) (INEC,2012b) para una precisa identificación de los perfiles laborales demandados. Además, se emplean métodos de aprendizaje automático, utilizando *Extreme Gradient Boosting* (XGBoost), para garantizar la completitud en variables críticas en el análisis como son: el nivel de educación y experiencia laboral requeridos.

Tras ejecutar el proceso de *web scraping* durante 8 semanas dentro del periodo del 18 de junio al 4 de septiembre de 2023, se recolectaron 70.700 anuncios de empleo, correspondientes a 194.814 vacantes disponibles. Sin embargo, para obtener una visión integral del mercado laboral ecuatoriano, se creó una base de datos con anuncios laborales únicos, basándose en criterios como el nombre de la empresa, cargo solicitado y descripción de la oferta. Esto resultó en un total de 40.939 anuncios y 93.586 vacantes. El análisis de la base de datos depurada reveló que la mayoría de las vacantes se ubican en los sectores de servicios, comercio y manufactura. Los cargos más demandados son técnicos y profesionales de nivel medio, vendedores y, directores y gerentes. En cuanto a la educación y experiencia, la mayoría de las vacantes están dirigidas a candidatos con educación de bachillerato o grado universitario, y con experiencia de 1 a 3 años.

Por último, este trabajo representa un paso adelante en el análisis del mercado laboral ecuatoriano, aplicando técnicas de *web scraping* y *machine learning* para recopilar datos precisos sobre las vacantes. Estas herramientas no solo ayudan a cerrar la brecha entre la oferta y la demanda de trabajo, fomentando la estabilidad económica y social, sino que también ofrecen potencial para su aplicación en otros contextos, abriendo nuevas vías en la investigación del análisis laboral y del *Big Data*.

El documento se encuentra organizado de la siguiente manera: la primera sección ofrece una revisión bibliográfica que proporciona un contexto preliminar sobre el mercado laboral, así como sobre las experiencias previas en el uso del *Big Data* como herramienta para recabar información sobre la demanda laboral. La segunda sección describe con detalle el enfoque metodológico empleado para el desarrollo del software, incluyendo los procesos de recolección, depuración y estandarización de la información. La tercera sección presenta un panorama general de los resultados alcanzados mediante el uso del software desarrollado. Finalmente, la cuarta sección resume las conclusiones desprendidas del trabajo, así como las futuras innovaciones al software generado.

1. Revisión de literatura

1.1. Contextualización del mercado laboral

El mercado laboral es el espacio donde se encuentran la oferta, representada por los trabajadores, y la demanda, representada por los empleadores (Cárdenas, 2020). Según la teoría económica clásica, este mercado opera en condiciones de competencia perfecta, lo que significa que tanto los trabajadores como los empleadores tienen un conocimiento completo de las condiciones del mercado, como los criterios de empleo, las oportunidades de trabajo y el precio del trabajo (Borjas, 2016). Bajo esta premisa, los empleadores conocen los requisitos específicos de un determinado puesto de trabajo, lo que les permite encontrar trabajadores que cumplan con sus necesidades, mientras que los trabajadores buscan empleo según sus propias características o competencias y deciden la cantidad de mano

de obra que desean ofrecer. Esta interacción se entiende como un intercambio de actividades humanas que sirve de insumo para la producción de bienes y servicios (Cárdenas, 2020; Borjas, 2016).

El equilibrio óptimo en el mercado laboral, que se alcanza cuando hay pleno empleo y todas las vacantes están ocupadas, rara vez se cumple en la realidad. Esto se debe en gran parte a las imperfecciones del mercado, como la información incompleta, obstaculiza la movilidad laboral y dificulta que los trabajadores encuentren sectores que requieran sus habilidades específicas (Benítez et al., 2016). Estas imperfecciones generan fallas de mercado, manifestadas en la incapacidad de empleadores y trabajadores para obtener información completa sobre el precio y la calidad de lo que se intercambia, limitando su capacidad de tomar decisiones informadas. Como consecuencia, puede surgir una escasez de cualificaciones, tanto en la oferta como en la demanda, lo que a su vez contribuye a la segmentación del mercado laboral y, potencialmente, a condiciones de desempleo (Cárdenas, 2020; Benítez et al., 2016).

Este panorama es provocado en gran parte por los continuos saltos tecnológicos, los cuales generan una divergencia entre las habilidades requeridas por los empleadores y las habilidades que disponen los trabajadores (Benítez et al., 2016). No obstante, al mismo tiempo que el avance tecnológico acrecienta los desajustes dentro del mercado laboral, ha permitido la aparición de instituciones o procesos, como son los portales de empleo en línea que facilitan la conexión de información entre trabajadores y empleadores.

En resumen, los avances tecnológicos, como la automatización y la globalización, están transformando el mercado laboral, generando una necesidad de nuevas habilidades y servicios (Autor, 2015). Esta evolución ofrece oportunidades tanto para los buscadores de empleo como para los empleadores. De tal modo que, las herramientas tecnológicas y estrategias de decisión eficaces se vuelven fundamentales para navegar en este mercado laboral dinámico, permitiendo una adaptación fluida a sus continuos cambios (Frid-Nielsen, 2019).

1.2. El Big Data como herramienta estadística

El desarrollo tecnológico en el ámbito de la información y comunicación ha conducido a una acumulación sin precedentes de datos digitales (Baig et al., 2019). Instituciones, ya sean públicas o privadas, han capitalizado esta tendencia, empleando la gran cantidad de estos datos para guiar decisiones y moldear estrategias (Cárdenas, 2020). Esta vastedad de datos, obtenida de dispositivos interconectados y enriquecida con técnicas de análisis, se identifica como *Big Data*. Si bien este término alude principalmente al volumen de información, también se destaca por la rapidez en su generación y la diversidad de datos que engloba (Gontero y Menéndez, 2021).

El auge del *Big Data* ha complementado y enriquecido las fuentes tradicionales de información, facilitando políticas más precisas y dirigidas. Actualmente, se extrae información de múltiples fuentes, como páginas web, sensores, aplicaciones móviles, y transacciones financieras, entre otros (George y Haas, 2014). Sin embargo, la explotación eficiente de estos datos requiere algoritmos específicos para su estructuración y análisis. Técnicas como el *web crawling*[2] y *web scraping*[3] se han convertido en herramientas esenciales en esta tarea (Gontero y Menéndez, 2021; Cárdenas, 2020).

El *web scraping* es especialmente útil para la extracción de grandes volúmenes y estructura de datos de Internet sin necesidad de interacción humana (ten Bosch et al., 2018). Esta técnica ha demostrado su utilidad en diversas áreas temáticas. Por ejemplo, se ha utilizado para recopilar información de precios en línea de tiendas mayoristas y minoristas, lo que ha permitido la generación de índices de precios al consumidor (Oancea y Necula, 2019; Polidoro et al., 2015). También se ha empleado en el campo del agroturismo para obtener información relevante del área

[2] El *web crawling* está diseñado para navegar de manera organizada y estructurada a través de un sitio web con el fin de catalogar su contenido (Nigam y Biswas, 2021).
[3] El *web scraping* se enfoca en la recopilación selectiva de información específica de las páginas web (Nigam y Biswas, 2021).

(Barcaroli et al., 2016), así como para recopilar datos sobre prácticas de sostenibilidad medioambiental, económica y sociocultural (Marchi et al., 2021; Sozzi, 2017). Además, se ha aplicado en la generación de estadísticas de empleo basadas en portales en línea (Cárdenas, 2020; Benítez et al., 2016; Cárdenas et al., 2015).

1.3. Web scraping y mercado laboral

En los últimos años, la disponibilidad de información en Internet ha aumentado considerablemente, un fenómeno destacado por Cárdenas (2020). Este crecimiento ha impactado significativamente en diversos sectores, incluido el mercado laboral. Este último se ha transformado por el uso de portales de empleo en línea, que actúan como intermediarios laborales y aglutinan una gran cantidad de datos sobre ofertas de empleo, características de las vacantes y requisitos de cualificación para cada puesto.

Según Maurer y Liu (2007), los empleadores están recurriendo cada vez más a estos portales para cubrir sus necesidades de personal, evidenciado por millones de vacantes y currículos disponibles en línea. Kässi y Lehdonvirta (2018) también señalan un incremento notable en el número de vacantes publicadas diariamente a nivel mundial. Esta tendencia resalta la importancia creciente de los portales de empleo en la dinámica del mercado laboral actual.

El aprovechamiento de la información contenida en los portales de empleo resulta un aspecto clave para generar estadísticas actualizadas y detalladas sobre el mercado laboral, lo cual a su vez es esencial en la formulación de políticas públicas eficaces (Carrillo y Vásconez, 2019). Esta práctica permite un seguimiento en tiempo real de la demanda laboral de las empresas y oferta de mano de obra de los trabajadores, identificar sus necesidades, minimizar la brecha informativa entre trabajadores y empleadores, y abordar el desajuste de habilidades en el mercado. Además, ofrece una visión clara sobre la evolución y el cambio en las cualificaciones laborales requeridas por las empresas (OECD, 2022; Cárdenas, 2020; Carrillo y Vásconez, 2019; Cárdenas et al., 2015).

La recolección de la información puede darse a través de convenios o por medio de algoritmos especializados de *web crawling* y *web scraping*. Estos últimos permiten la navegación eficaz por los componentes de las páginas web para la extracción organizada de información. Así se logra caracterizar con precisión la demanda de empleo, incluyendo información detallada de las empresas, y, según el contexto, también la oferta laboral, es decir, datos de los buscadores de empleo (Gontero y Menéndez, 2021).

Aunque las técnicas de *web scraping* son una herramienta útil y de menor costo para recopilar información del mercado laboral, no puede sustituir las metodologías tradicionales de recolección de datos, como encuestas de hogares, registros administrativos o censos. Dado que estas operaciones estadísticas están diseñadas para capturar datos representativos de ciertos universos de estudio (Cárdenas et al., 2015). Según Gontero y Menéndez (2021), una limitación significativa de los datos obtenidos mediante *web scraping* es su representatividad. Es común que ciertos sectores, tamaños de empresas o tipos de empleo estén más presentes en portales de empleo en línea, mientras que otros estén subrepresentados.

Otro aspecto a contemplar es la posible duplicidad de anuncios de empleo en diferentes portales, o incluso dentro del mismo portal, cuando las empresas publican en múltiples sitios para cubrir sus vacantes. Esta práctica puede llevar a una percepción errónea de la cantidad de empleos disponibles. Adicionalmente, la naturaleza dinámica de los portales de empleo, donde los anuncios aparecen y desaparecen constantemente, afecta la precisión de los datos recopilados. Por último, el procesamiento de la información capturada podría introducir sesgos involuntarios al momento de aplicar técnicas especializadas de procesamiento de texto para la obtención de datos adecuados para el análisis. A pesar de que el uso extensivo de datos ofrece ventajas significativas para el desarrollo de sistemas de información, es crucial tener consciencia de estas limitaciones y abordarlas adecuadamente (Gontero y Menéndez, 2021).

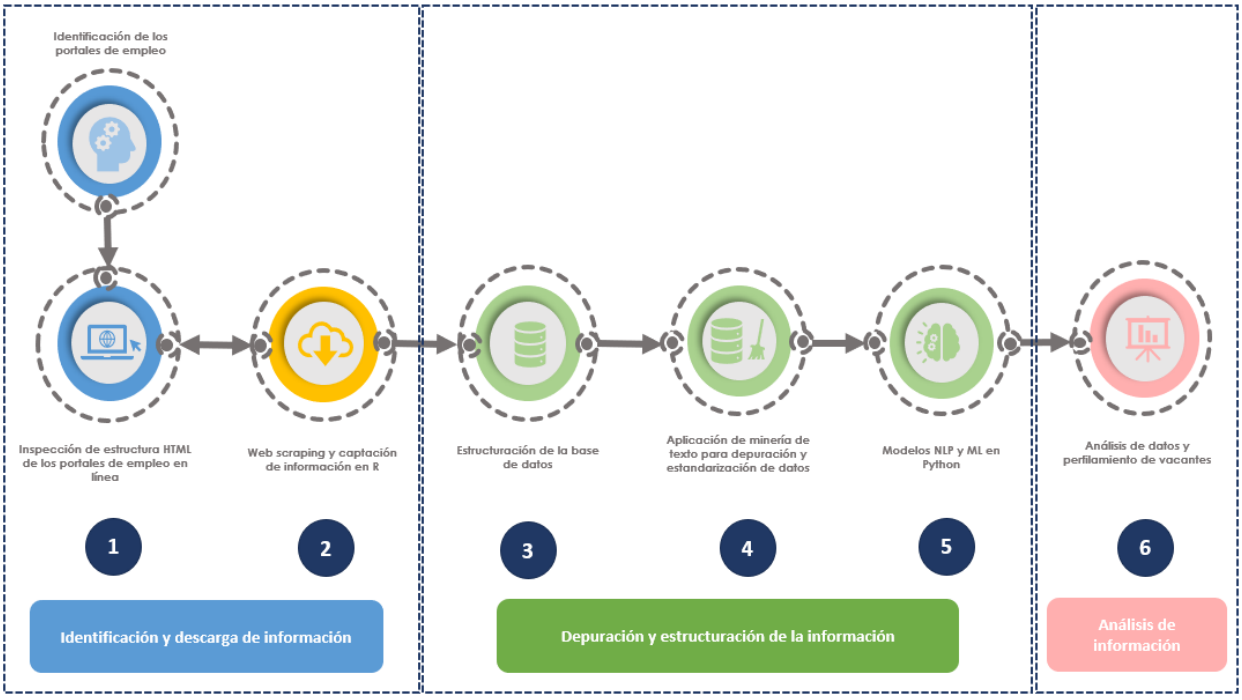
A nivel internacional, diversos países han demostrado un uso innovador de la información obtenida de portales de empleo en línea para analizar tendencias laborales. Por ejemplo, en Australia, el Gobierno ha desarrollado el *Internet Vacancy Index*, un índice que capitaliza los anuncios en bolsas de trabajo para medir las vacantes laborales (Australian Government, 2023). En Estados Unidos, se destaca el *Help-Wanted Index*, creado por la *Conference Board* y *Burning Glass Technologies*, que se basa en el seguimiento de anuncios laborales (Gontero y Menéndez, 2021). Por su parte, Nueva Zelanda ha implementado el *All Vacancy Index*, que utiliza datos de cuatro sitios web y ofrece una clasificación detallada de las vacantes por industria, ocupación y competencia (Gontero y Menéndez, 2021).

En Latinoamérica, el uso de portales de empleo en línea ha impulsado varios proyectos analíticos significativos. En Uruguay, se ha realizado un estudio pionero sobre la dinámica de las cualificaciones utilizando estas herramientas digitales. De forma más amplia, la CEPAL, en 2019, llevó a cabo un proyecto exhaustivo, recopilando y analizando las características de las vacantes de 8 portales en 33 países latinoamericanos. Este estudio se centró en evaluar las tendencias actuales del mercado laboral (Gontero y Menéndez, 2021). Por otro lado, el Banco Interamericano de Desarrollo (BID) y la Asociación de Economía de América Latina y el Caribe (LACEA) colaboraron en 2018 en un análisis de las vacantes de 12 portales en cinco países, con un enfoque particular en los requerimientos educativos y de experiencia (Gontero y Menéndez, 2021). Adicionalmente, investigadores como Benítez et al. (2016) han aplicado técnicas de *web scraping* para extraer datos del portal Computrabajo en Ecuador, perfilando de manera precisa las vacantes laborales disponibles. De manera similar, Cárdenas et al. (2015) en Colombia propusieron un enfoque innovador para generar estadísticas de demanda laboral, basándose también en técnicas de *web scraping*.

2. Implementación metodológica

La implementación metodológica seguida en este trabajo se basa en el desarrollo de un software especializado en la generación de estadísticas de vacantes, integrando algoritmos de *web scraping*, *web crawling*, minería de texto, técnicas de *natural language processing* (NLP) y modelado a través de *machine learning*. El software desarrollado abarca un proceso completo que va desde la identificación y recolección de datos hasta su depuración, estructuración y el análisis exhaustivo de la información procesada, tal como se muestra en la Figura 1.

Figura 1. Diseño de software para la generación de estadísticas de vacantes



En los siguientes apartados, se detallan las etapas específicas que constituyen la experiencia de desarrollo del software dirigido a la generación de estadísticas de vacantes.

2.1. Identificación y descarga de información

Las plataformas de empleo en línea se han convertido en repositorios extensos de datos sobre las vacantes ofertadas por las empresas. Sin embargo, la utilización de estos datos para generar estadísticas laborales representa un desafío significativo. Según Cárdenas (2020) y Cárdenas et al. (2015), la tarea de recolectar y sistematizar esta información para análisis es complejizada por la creciente cantidad de anuncios de trabajo y el aumento rápido de portales especializados. Estos sitios web varían considerablemente en términos de diseño, lenguaje de programación, estructura y presentación de datos, lo que a menudo resulta en la duplicidad de información entre diferentes plataformas.

Ante este escenario, se vuelve necesario establecer una metodología robusta para identificar los portales de empleo más pertinentes. Esta metodología, como sugiere Cárdenas (2020), debería basarse en criterios bien definidos, incluyendo el volumen de vacantes disponibles, la calidad y la organización del portal, y el número de usuarios activos. Una vez seleccionados los portales relevantes, se requiere de una estrategia meticulosa para la captura y el almacenamiento efectivo de la información, garantizando así la generación de estadísticas laborales coherentes y precisas.

Según la metodología propuesta, se han elegido tres plataformas en Ecuador para el estudio: Multitrabajos (2023) y Computrabajo (2023), que son portales privados, y Encuentraempleo (2023), un sitio público. Estos portales han sido identificados, según Benítez et al. (2016), como los más concurridos en el país, acumulando una cantidad significativa de anuncios de empleo publicados periódicamente: alrededor de 6.000 en Multitrabajos, 2.500 en Computrabajo y cerca de 1.000 en Encuentraempleo. Estos números reflejan el volumen de vacantes disponibles, uno de los criterios clave en la metodología propuesta. Además, estos sitios presentan una ventaja adicional: su información se encuentra en un formato semi-estructurado[4], lo cual facilita la captura y organización de los datos para su posterior análisis. Esta característica asegura una recolección de datos eficiente y precisa, en línea con los desafíos y soluciones identificados en estudios previos.

Siguiendo la identificación de los portales de empleo a ser analizados, el siguiente paso corresponde a la captura y sistematización de la información. En ese contexto a partir de la estrategia descrita por Cárdenas (2020) y Cárdenas et al. (2015), se combinan técnicas de *web crawling* y *web scraping*, utilizando algoritmos que navegan a través de las distintas estructuras de las páginas web y extraen datos de forma automatizada. En particular, el algoritmo[5] implementado simula la interacción humana al visitar varios portales de empleo, efectuando *crawling* para navegar a través del portal y ubicar la información deseada, y *scraping* para su recolección y posterior almacenamiento en bases de datos semiestructuradas, donde la definición de los campos es más flexible. Este enfoque explota la arquitectura de los sitios web y el lenguaje HTML, utilizando los nodos de programación conocidos como *Xpaths* para identificar de forma precisa elementos como color, forma y contenido, lo que incluye etiquetas, encabezados, entre otros.

Complementando esta técnica, se ha implementado una rutina de recopilación de datos que se ejecuta semanalmente en cada uno de los portales[6]. Este procedimiento se ha llevado a cabo consistentemente durante los meses de junio a septiembre de 2023, lo que garantizó la captura de una amplia gama de anuncios de empleo y mantiene la base de

[4] Las diversas fuentes de información generan datos en formatos variados, que pueden ser estructurados, semiestructurados o no estructurados, como describen Cárdenas et al. (2015). Los datos estructurados se distinguen por su organización en campos definidos, consistentes y en un orden específico. Por otro lado, los datos semiestructurados, aunque carecen de una estructura de campos rígida, contienen elementos como etiquetas y encabezados que facilitan la identificación y el manejo de la información contenida. En contraste, los datos no estructurados no poseen una organización fija ni elementos distintivos que permitan discernir de manera directa los componentes de la información, representando así un desafío mayor para su procesamiento y análisis detallado.

[5] Para cada uno de los portales de empleo, se diseñó un algoritmo distinto. Estas diferencias en los códigos se debieron principalmente a la necesidad de adaptarse a la navegación y extracción de datos en cada portal web, considerando que cada uno posee una estructura única.

[6] El tiempo de descarga varía entre los diferentes portales, influenciado primordialmente por la estructura de cada sitio web. Específicamente, para los portales Encuentraempleo y Computrabajo, el proceso de descarga toma aproximadamente dos horas. Por otro lado, para Multitrabajos, el tiempo requerido para completar una descarga se extendió a casi a un día completo.

datos actualizada con las últimas ofertas laborales. La periodicidad de esta recopilación no solo refleja las tendencias actuales del mercado laboral, sino que también asegura que los análisis estadísticos derivados sean pertinentes y reflejen el cambiante contexto laboral ecuatoriano.

2.2. Depuración y estructuración de la información

2.2.1. Limpieza de información

Una vez se han extraído y almacenado los anuncios de empleo en las bases de datos, es fundamental aplicar distintos pasos en cada una para garantizar la coherencia y consistencia de la información. Siguiendo las metodologías utilizadas por Benítez et al. (2016) y Cárdenas et al. (2015), dentro de los pasos relevantes de depuración se encuentran: 1) la eliminación de anuncios duplicados, 2) la limpieza y la normalización de la base de datos, y 3) la generación de variables significativas aprovechando técnicas de minería de texto.

La eliminación de anuncios duplicados es esencial, ya que pueden darse casos en los que anuncios se encuentren publicados múltiples veces en un mismo portal, lo cual podría inflar artificialmente las cifras de anuncios de empleo y de vacantes disponibles (Gontero y Menéndez, 2021). Para erradicar este problema, se ha establecido la eliminación de casos duplicados de acuerdo a criterios de coincidencia que incluyen la empresa, el cargo solicitado, la descripción detallada del anuncio y la fecha de publicación.

La limpieza de la base de datos se basa en la eliminación de elementos irrelevantes, como es el caso de caracteres especiales, signos de puntuación y *stop word*[7]*s*, que no aportan al análisis. Lo que combinado con la transformación de los textos de las bases de datos de mayúsculas a minúsculas aseguran la uniformidad en la información.

A continuación, para enriquecer a la base de datos, se implementan estrategias de minería de texto[8]. Esta etapa implica la creación de variables clave y la extracción de información de campos preexistentes mediante el uso de patrones textuales y expresiones regulares encontrados en las descripciones detalladas de los anuncios. Este proceso, basado en las metodologías de Benítez et al. (2016) y Cárdenas et al. (2020), permite obtener datos estructurados como son: el tipo de contrato, la jornada laboral, el salario, la ubicación geográfica (provincia y cantón), los requisitos educativos, la experiencia requerida, la disposición a viajar, la inclusión de personas con discapacidad, si se requiere licencia de conducir y el número de vacantes disponibles.

La última fase del proceso de depuración, consiste en consolidar las bases de datos depuradas en una única base de datos maestra. A la cual, se realiza una eliminación adicional de duplicados, considerando la posibilidad de que los empleadores publiquen el mismo anuncio en varios portales. Se mantienen los mismos criterios de coincidencia (cargo solicitado, la descripción detallada del anuncio y la fecha de publicación) para asegurar la precisión en la representación de las ofertas laborales disponibles en un período específico, como es la misma semana de descarga de información.

2.2.2. Estandarización de la información

El proceso de estandarización de la información implementado en este estudio se centró en la definición y creación de cuatro variables clave. Dos de estas variables se dedicaron a estandarizar la información geográfica de los anuncios de empleo, tomando como referencia la provincia y el cantón especificados en los mismos. La comparación de estas variables con la DPA de Ecuador (INEC, 2022) facilitó la precisión en las descripciones geográficas y una codificación efectiva, empleando patrones textuales para la clasificación.

Además, se desarrollaron modelos para clasificar la rama de actividad y el cargo ocupacional, recurriendo a catálogos estandarizados como la CIIU.4.0 (INEC, 2012a) y CIUO.08 (INEC, 2012b). Este enfoque innovador mejora las metodologías de trabajos anteriores, tales como los de Cárdenas (2020) y Benítez et al. (2016), al integrar técnicas de

[7] Palabras que no aportan a la comprensión de un texto como preposiciones, conjunciones o artículos.
[8] La minería de texto es definida como el proceso de analizar texto para extraer información útil para determinados fines (Valdez-Almada et al., 2017).

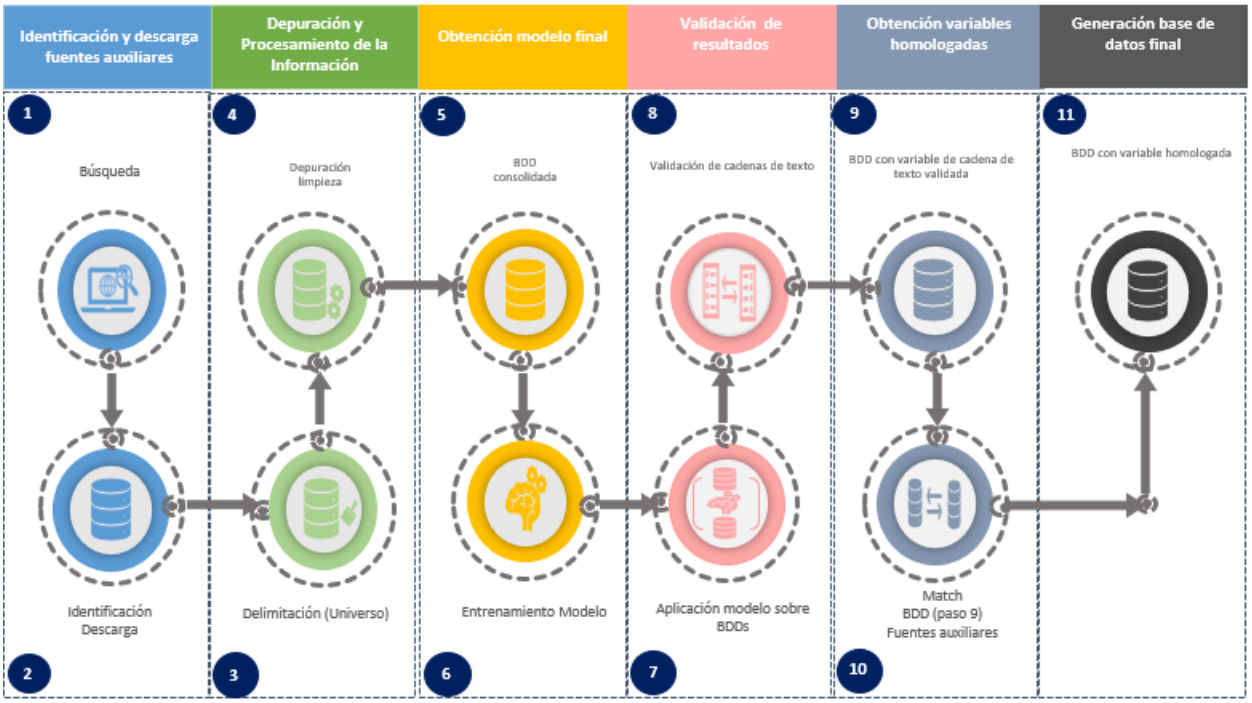
machine learning y modelos de similitud de texto basados en transformadores. Esta combinación ha permitido alcanzar un nivel de precisión y detalle sin precedentes en la clasificación de la información.

La aplicación de modelos de similitud de texto[9], es una técnica avanzada del NLP, la cual busca evaluar la correspondencia entre dos segmentos de texto mediante una métrica específica. En este contexto, los modelos basados en transformadores (Hirschberg y Manning, 2015; Khurana et al., 2023), gracias a su capacidad de retención de contexto, han sido fundamentales para clasificar eficazmente la información (Vaswani et al., 2017).

Estos modelos preentrenados con grandes conjuntos de datos provenientes de múltiples dominios (Devlin et al., 2019; Min et al., 2021) pueden requerir ajustes para mejorar su precisión en tareas o dominios especializados. Por lo que, para abordar estas limitaciones, se emplea un proceso de afinamiento (*fine-tuning*) (Raffel et al., 2020), que implica entrenar de nuevo el modelo con un conjunto de datos específicos del dominio de interés. Este enfoque comienza con la selección de fuentes de información auxiliares y la preparación de un conjunto de datos de entrenamiento adecuado para el *fine-tuning*. Posteriormente, el modelo ajustado se utiliza para comparar las descripciones de los anuncios de empleo con la información auxiliar, validando la similitud entre las descripciones obtenidas y los datos de referencia. Este procedimiento facilita la correcta clasificación de las ofertas de empleo según la rama de actividad (CIIU.4.0) y el cargo ocupacional (CIIU.08), siguiendo una secuencia lógica que comienza con la categorización de la rama de actividad.

La metodología desarrollada establece un marco coherente para la estandarización y clasificación de datos laborales, representando un avance significativo en la precisión y aplicabilidad de la información recopilada para análisis del mercado laboral, el proceso es resumido en la Figura 2.

Figura 2. Proceso de obtención de variables homologadas



[9] En el contexto del desarrollo del software, se realizaron pruebas con modelos preentrenados de clasificación de texto basados en *Bidirectional Encoder Representations from Transformers* (BERT) y redes recurrentes de tipo Long Short-Term Memory (LSTM). Sin embargo, la precisión de la clasificación no fue la óptima, lo cual se atribuye a la elección de las encuestas de empleo como fuente de datos para el entrenamiento de los modelos. Esta fuente de información presentó una estructura de los textos de rama de actividad y cargo ocupacional distinta a la presentada en las descripciones de anuncios de empleo, lo que resultó en discrepancias significativas que afectaron la precisión de la clasificación.

Generación de rama de actividad a partir del CIIU.4.0

La obtención de información correspondiente a la rama de actividad a partir del CIIU.4.0 se llevó a cabo principalmente a través de fuentes auxiliares provenientes de institucionales públicas, en particular, el Servicio de Rentas Internas (SRI) y la Superintendencia de Compañías, Valores y Seguros (SuperCias). La información utilizada es de naturaleza administrativa y se conforma por registros detallados de sociedades y personas naturales, como son denominaciones de la empresa y ubicación geográfica, entre otros. La Tabla 1 resume la información disponible.

Tabla 1. Fuentes de información disponibles

Fuente de Información	Nombre del registro	N° bases de datos	N° de registros	N° de columnas
SRI	Registro Único de Contribuyentes*	24	307.495**	20***
	Catastro de grandes contribuyentes	1	500	8
SuperCias	Directorio de Compañías	1	199.135	23

Nota:
* El registro administrativo Registro Único de Contribuyentes presenta información de todos los contribuyentes (empresas y personas naturales) por varias características tales como: rama de actividad, estado del contribuyente, entre otros. Sin embargo, esta información no se encuentra consolidada en una única base de datos nacional; sino más bien, se distribuye en 24 bases provinciales.
** Número de registros promedio
*** Número de columnas promedio

El procesamiento de la información comienza definiendo el universo de estudio, el cual está formado por las empresas activas inscritas en la SuperCias, así como sociedades y un pequeño grupo de personas naturales registradas en el SRI. Se analiza principalmente la razón social o nombre comercial de los contribuyentes. Para mejorar la precisión del análisis, en el caso de las personas naturales, se han seleccionado solamente aquellos registros que incluyen nombres comerciales o denominaciones auxiliares, conocidos como nombre de fantasía. Esto ayuda a evitar confusiones al diferenciar entre individuos que, a pesar de tener roles o actividades económicas distintas, podrían compartir el mismo nombre, lo cual podría generar incertidumbre en la selección.

Adicionalmente, se conservan datos que ofrecen información sobre la rama de actividad y la razón social de los contribuyentes. También se eliminan los valores perdidos y se mantienen aquellos registros que contengan códigos CIIU estandarizados[10]. Este enfoque asegura que el análisis sea más claro y evita posibles ambigüedades al seleccionar los datos para su estudio.

La base final es una consolidación de la información disponible que finaliza con 228.641 registros distribuidos en 2 columnas. La primera columna compila todas las posibles denominaciones de sociedades y nombres de fantasía de personas naturales, tras haber sido depurada de caracteres especiales y espacios excesivos. La segunda columna alberga la actividad económica recodificada, transformando sus valores de alfanuméricos a secuencias numéricas.

Tras construir las bases de datos de entrenamiento, el siguiente paso consistió en definir el modelo a implementar. Este se concreta mediante el entrenamiento del algoritmo *all-mpnet-base-v2*, un algoritmo altamente reconocido por su versatilidad y eficiente rendimiento, con una demanda relativamente baja de recursos computacionales, según lo destacan el *Massive Text Embedding Benchmark* (Muennighoff et al., 2023) [11] y el ranking de Reimers y Gurevych (2019)[12].

[10] La rama de actividad obtenida desde SRI presenta ramas adicionales en comparación a la Clasificación Internacional Uniforme (CIIU) debido a que esta institución utiliza su propia adaptación del CIIU.
[11] Las métricas obtenidas por los distintos algoritmos, así como su posición dentro de este ranking pueden ser encontrados en el siguiente URL: https://huggingface.co/spaces/mteb/leaderboard.
[12] Las métricas obtenidas por los distintos algoritmos, así como su posición dentro de este ranking pueden ser encontrados en el siguiente URL: https://www.sbert.net/docs/pretrained_models.html.

En cuanto a la selección de hiperparámetros para optimizar el modelo, se asignan valores a cuatro elementos clave: *epochs, batch size, warm-up steps* y *loss function*. Los valores seleccionados buscan maximizar el uso de los recursos computacionales disponibles, estableciéndose en: *epochs* (10), *batch size* (16) y *warm-up steps* (10.454).

Por otro lado, se opta por la función de pérdida *BatchHardSoftMarginTripletLoss* (Hermans et al., 2017), una decisión resultante de un exhaustivo proceso de comparación sistemática. Este proceso implica crear un modelo para cada función de pérdida potencial y aplicar estos modelos sobre los conjuntos de datos del catastro de grandes contribuyentes del SRI y el directorio de compañías de la SuperCias. La evaluación de los resultados se realiza manualmente, mediante la comparación de las respuestas obtenidas y la generación de una métrica de precisión análoga al *accuracy*. Los resultados de este proceso se presentan en la Tabla 2.

Tabla 2. Resultado contraste *loss function* potenciales

Función de pérdida	Hiperparámetros	accuracy*
BatchAllTripletLoss	**epoch:** 3 **warm up steps:** 3.136 **batch size:** 16	66,00%
BatchHardSoftMarginTripletLoss		85,80%
BatchHardTripletLoss		83,60%
BatchSemiHardTripletLoss		0,20%

Nota:
***** El cálculo de este *accuracy* personalizado es igual al total de textos similares (predicción correcta) sobre el n° total de registros.

Una vez obtenido el modelo final, se procede a aplicarlo a la base de datos de ofertas de empleo recopiladas de la web (*query)*[13], enfocándose específicamente en la variable nombre de la empresa. Para la comparación, se utiliza nuevamente la base de entrenamiento previamente generada (*corpus)*[14]. Esta base de datos se selecciona debido a su exhaustividad, ya que abarca un amplio número de sociedades y personas naturales. Además, destaca por su diversidad, incluyendo no solo denominaciones legales sino también nombres alternativos o de fantasía.

En esta fase, se realiza una evaluación minuciosa de cuatro métodos distintos para aplicar el modelo. El objetivo es determinar cuál de estos enfoques ofrece los resultados más efectivos y precisos. Este análisis crítico permite asegurar que la implementación del modelo sea la más adecuada para las necesidades y características específicas de la base de datos de ofertas de empleo.

El primer método se centra en la implementación básica del modelo, aplicándolo directamente sobre las bases de datos *query* y de *corpus* para luego validar los resultados obtenidos. El segundo método introduce la estrategia de *retrieve & rerank*[15] (Geigle et al. ,2022), durante la fase de selección de cadenas de texto, optimizando así la eficiencia en la recuperación de información relevante. El tercer método consiste en una validación y aplicación iterativa del modelo, dividiendo la base de datos de corpus en múltiples segmentos (cuatro en total). A partir de la segunda iteración, la base de datos de *query* se reduce a medida que se han validado los registros en iteraciones previas. Finalmente, el cuarto método combina las técnicas de *retrieve & rerank* con el proceso iterativo de aplicación y validación del modelo, ofreciendo un enfoque comprensivo. La Tabla 3 proporciona un resumen detallado de los métodos descritos.

[13] La base *query* es el conjunto de datos que contiene los fragmentos de texto que se desea emparejar con información auxiliar.
[14] La base corpus es la base de cadenas de texto, desde donde el modelo extraerá las distintas opciones para emparejar con la base *query*.
[15] La metodología consiste en la selección de resultados en dos etapas. En la primera etapa *retrieve* se utiliza un algoritmo (*bi-encoder*) que realiza una primera selección de resultados que pasan a la etapa de *rerank* donde, a partir de un segundo algoritmo (*cross-encoder*), los resultados son reevaluados.

Tabla 3. Métodos de aplicación del modelo

Nombre	Características
Método 1	**Base *query*:** Única base **Base *corpus*:** Única base **Selección de cadenas de texto:** A partir del modelo
Método 2	**Base *query*:** Única base **Base *corpus*:** Única base **Selección de cadenas de texto:** Aplicación de estrategia *retrieve & rerank*
Método 3	**Base *query*:** múltiples bases **Base *corpus*:** múltiples bases **Selección de cadenas de texto:** A partir del modelo
Método 4	**Base *query*:** múltiples bases **Base *corpus*:** múltiples bases **Estrategia de adicionales:** Aplicación de estrategia *retrieve & rerank*

En el proceso de evaluación de los cuatro modelos propuestos, se lleva a cabo un exhaustivo análisis de los resultados obtenidos por cada uno. Este análisis consistió en una revisión minuciosa de la coincidencia entre la cadena de texto buscada y la sugerida por el modelo; siendo considerada como predicción correcta al registro con cadenas de texto similares. La precisión final para cada modelo se obtiene del total de predicciones correctas identificadas. Para este propósito, se emplea la similitud de coseno[16] como métrica principal. Esta métrica es asignada por el modelo y representa el grado de similitud entre cadenas de texto, con valores en el rango de-1 y 1 (los valores cercanos a uno son preferidos).

Es importante señalar que, durante la validación de las predicciones, se consideran las cadenas de texto similares en todos los rangos de la similitud de coseno, asegurando así una evaluación integral y detallada de la capacidad predictiva de cada modelo en distintos niveles de similitud textual. Los resultados se presentan en la Tabla 4 tanto de forma absoluta como relativa. El modelo sugiere tres opciones posibles por cada cadena de texto buscada; sin embargo, se evalúa solo la primera opción que a su vez posee la similitud de coseno más alta.

Tabla 4. Resultados comparación opciones de aplicación del modelo

Estrategia de val	Método 1		Método 2		Método 3		Método 4	
Rangos	Absoluto	Relativo	Absoluto	Relativo	Absoluto	Relativo	Absoluto	Relativo
1 - 0.90	2.629	79,93%	2.632	79,64%	2.402	72,22%	2.438	71,08%
0.89 -0.80	200	6,08%	203	6,14%	222	6,67%	234	6,82%
0.79 – 0.70	158	4,80%	161	4,87%	203	6,10%	208	6,06%
0.69 -0.60	162	4,93%	164	4,96%	198	5,95%	225	6,56%
0.59 – 0.50	96	2,92%	99	3,00%	162	4,87%	174	5,07%
0.49 – 0.40	29	0,88%	31	0,94%	86	2,59%	95	2,77%
0.39 – 0.30	14	0,43%	14	0,42%	47	1,41%	48	1,40%
0.29 – 0.0	1	0,03%	1	0,03%	6	0,18%	8	0,23%
Total	3.289	100,00%	3.305	100,00%	3.326	100,00%	3.430	100,00%

Nota:
Se presenta únicamente los resultados de la opción con la métrica de coseno más alta.

[16] La similitud entre fragmentos de texto puede evaluarse mediante diversas métricas, incluyendo la correlación de Spearman o Pearson, las distancias Euclidianas o de Manhattan, y la similitud de coseno, entre otras. Sin embargo, la similitud de coseno se destaca como la métrica preferida para la evaluación. En este trabajo, se utilizó la biblioteca de Python *Sentence-Transformers*, que facilita la evaluación de similitud en gran escala, siendo la similitud de coseno la métrica estándar empleada.

Al comparar los resultados absolutos y relativos, se aprecia que a mayor rango de similitud de coseno mayor número de predicciones correctas obtenidas. Asimismo, la aplicación iterativa del modelo muestra una mejora en la precisión general del mismo (totales); como es el caso del total (3.326) generado en el método 3 (proceso iterativo) con respecto al total (3.289) obtenido desde el método 1 (modelo base). Adicionalmente, la inclusión de la estrategia *retrieve & rerank* muestra un efecto positivo que se evidencia al comparar la precisión (3.430) en el método 4 (proceso iterativo - *retrieve & rerank)* con respeto a la precisión (3.326) del método 3 (proceso iterativo).

Al combinar ambas estrategias (proceso iterativo- *retrieve & rerank)* se obtiene el modelo con la mejor precisión (3.430) de las opciones comparadas; por lo cual, se determina al método 4 (proceso iterativo- *retrieve & rerank)* como la mejor aplicación del modelo. Con la elección del mejor método de aplicación se realiza un segundo proceso de validación de resultados que permita identificar todos los posibles emparejamientos. Este nuevo proceso amplía las validaciones realizadas previamente, ya que considera todas las opciones de cadenas de texto arrojadas por el modelo y no solo aquella con mayor similitud de coseno. Igualmente, se realiza una segunda verificación de la similitud entre cadenas de texto, tomando como referencia una nueva métrica, la cual se denomina *cross-score*, que es asignada durante la aplicación de la estrategia *retrieve & rerank*.

No obstante, no todos los nombres de empresas en la base de ofertas web encontraron una coincidencia en las fuentes de información auxiliares de manera automática[17]. En respuesta a esta situación, se llevó a cabo una serie de procesos de emparejamiento manual, que son posibles, por la riqueza de información que se posee para la rama de actividad, este emparejamiento consistió en tres acciones principalmente. La primera determina grupos de empresas que pudieran encontrarse en una misma rama de actividad económica, como pueden ser unidades educativas, hospitales o instituciones públicas. La siguiente acción consistió en eliminar caracteres o cadenas de texto generales que pudieran afectar el emparejamiento como pueden ser terminaciones (cia, ltda, etc.), ubicaciones (ecuador, pichincha) e incluso espacios entre caracteres. La tercera realiza una búsqueda manual específica a aquellas empresas que concentren un gran número de anuncios de empleo. Los registros que a pesar de todas las opciones de emparejamiento (manual y automático) no pudieron ser identificados se les asigna un código numérico (99999) que hace alusión a valor perdido.

Generación de rama de actividad a partir del CIUO.08

Posterior a la creación de la variable de rama de actividad en la base de datos de anuncios de empleo, el paso siguiente es la generación del cargo ocupacional. Es importante reconocer que este proceso enfrenta limitaciones significativas, especialmente en cuanto a la disponibilidad de información en el contexto ecuatoriano, lo cual impide la validación de los resultados obtenidos. Estas circunstancias hacen que la modelización del cargo ocupacional difiera notablemente de la realizada para la rama de actividad. Sin embargo, se decide aplicar las mejores prácticas identificadas durante la modelización de la rama de actividad.

Para el análisis del cargo ocupacional, se identifica como fuente principal el Catálogo Nacional de Cualificaciones Profesionales del Ministerio de Trabajo (MDT). No obstante, la información del MDT no se encuentra en formato de base de datos, sino en documentos en formato PDF que describen detalladamente las actividades específicas y generales de diversas ocupaciones. Ante la naturaleza no estructurada de estos datos, se implementa un proceso iterativo de extracción de información.

Este proceso utiliza patrones y expresiones regulares para identificar y extraer con precisión tanto la descripción del cargo como las actividades asociadas a cada ocupación. Así, se convierte la información inicialmente no estructurada en dos bases de datos diferenciadas: 1) base de datos con información sobre la descripción del cargo ocupacional y de actividades generales; 2) base de datos correspondiente a la descripción del cargo ocupacional y de actividades específicas.

[17] Se debe precisar que el modelo arroja una coincidencia para cada registro presente en la base *query*; sin embargo, al validar los resultados puede existir la posibilidad que la opción sugerida desde el modelo no tenga relación con la opción buscada.

Paralelamente, se crea una tercera base de datos utilizando la clasificación CIUO.08. Se procesa la descripción de los cargos de manera similar a lo hecho con la rama de actividad, eliminando caracteres especiales y espacios excesivos. De este proceso surge una base de datos adicional con dos columnas, como son el nombre del cargo y el código de cuatro dígitos correspondiente a la clasificación CIUO.08.

Después de crear las bases de datos, se inicia la fase de modelización del cargo ocupacional. Sin embargo, debido a la notable disparidad en la calidad de la información en comparación con la rama de actividad, se realizan ajustes en la estrategia de modelización. Para abordar estas diferencias, se elige utilizar el algoritmo *Sentence-T5-Large*[18], especializado en tareas de similitud de texto, aunque requiere de una mayor capacidad computacional. Este enfoque resulta en la creación de tres modelos, cada uno asociado a las diferentes bases de datos previamente elaboradas.

En cuanto a los hiperparámetros, se conserva el valor de *epoch* utilizado para la rama de actividad, mientras que se reduce el tamaño del *batch size* a 8 para optimizar los recursos computacionales. Los *warm up steps* se ajustan de manera diferente según la base de datos a la que se aplique.

Para las funciones de pérdida, se utiliza *MultipleNegativesRankingLoss* (Henderson et al., 2017) en los modelos derivados de las bases de datos del MDT, dada su eficacia en la generación de resultados. Por otro lado, para la base de datos derivada de la clasificación CIUO.08, se emplea la misma función de pérdida que en el modelo de la rama de actividad (*BatchHardSoftMarginTripletLoss*). Esta diferenciación en las funciones de pérdida se adapta a las características específicas de cada conjunto de datos, asegurando así la eficiencia del entrenamiento en cada contexto.

Posteriormente, se aplica el modelo a los puestos de trabajo listados en las ofertas de empleo descargadas y a las ocupaciones descritas en el clasificador CIUO.08. Al igual que en la rama de actividad, se utiliza la estrategia de *retrieve & rerank* para la selección de resultados.

Debido a la imposibilidad de validar los resultados de manera externa, se realiza un proceso comparativo entre los resultados de cada modelo. En la primera etapa, se define la mejor cadena de texto generada por cada modelo, considerando tanto la similitud de coseno como el *cross-score*, aunque se da mayor peso a esta última métrica. En la segunda y última etapa, se elige la mejor cadena de texto de los resultados obtenidos en la primera etapa, manteniendo el mismo método de evaluación. Por, ultimo, de manera similar a la rama de actividad, se asigna un código numérico (99999) a aquellos registros que no logran emparejarse, como un indicativo de valores perdidos.

2.2.3. Imputación de valores faltantes

Tras la construcción y estandarización de las variables pertinentes, el último enfoque metodológico se centró en la imputación de variables clave para el análisis del mercado laboral, en particular, el nivel de instrucción y la experiencia laboral. El proceso comenzó con la identificación de los tipos de datos faltantes, basándose en la clasificación propuesta por Rubin (1976), que los distingue en tres categorías: *Missing Completely at Random* (MCAR), *Missing at Random* (MAR) y *Not Missing at Random* (MNAR). En este trabajo, se determinó que los datos ausentes correspondían a la categoría MAR, lo que indica que su falta está relacionada con la información disponible en la base de datos. Este patrón es común en sitios web donde los empleadores detallan sus requisitos.

Con esta clasificación como punto de partida, se procedió a la etapa de imputación. Se adoptaron técnicas de *machine learning*, específicamente XGBoost. Siguiendo las metodologías sugeridas por INEC (2023), Rosati (2021) y Thamelo et al. (2021), la técnica XGBoost es conocida por su eficacia en la predicción y el manejo de valores faltantes, así como por su capacidad para adaptarse a diversas fuentes de datos (Rosati, 2021).

El proceso de imputación se dividió en dos fases. La primera consistió en un análisis exploratorio de todas las variables generadas en la base de datos. Se seleccionaron aquellas variables para la imputación que mostraban mayor consistencia y completitud, y que eran relevantes para el análisis de ofertas de empleo. Entre las variables elegidas se

[18] El algoritmo *Sentence-T5-Large* no solo es evaluado en los mismos rankings que el algoritmo *all-mpnet-base-v2* utilizado por la rama de actividad, sino que presenta un mejor desempeño en la tarea de similitud de texto.

incluyeron sexo, ubicación geográfica (provincia y cantón del anuncio), rama de actividad, cargo ocupacional, salarios, tipo de contrato, jornada laboral, modalidad, experiencia y nivel de instrucción. A continuación, mediante el análisis de correlación de Cramér[19], se identificaron las variables definitivas para los modelos, descartando aquellas con alta correlación, como la provincia del cargo solicitado.

La segunda etapa de imputación se centró en la aplicación de la técnica XGBoost para la clasificación de datos relacionados con el nivel de instrucción y la experiencia laboral. Para ello, se definieron hiperparámetros estandarizados para ambas variables, incluyendo métodos de búsqueda y optimización aleatoria[20], así como la implementación de la validación cruzada (k[21]=5). Estos pasos se tomaron con el objetivo de garantizar la consistencia y comparabilidad de los modelos, y al mismo tiempo, minimizar posibles divergencias técnicas. El desempeño de cada modelo se midió a través de la exactitud (*accuracy*)[22] en las observaciones del grupo de prueba para cada variable imputada.

Como se observa en la Tabla 5, los modelos implementados mostraron resultados satisfactorios en términos de exactitud, alineándose con las referencias de estudios previos como los de INEC (2023) y Rosati (2021), con métricas que superan el 70% de *accuracy*. Además, se notó una consistencia entre la exactitud del grupo de entrenamiento y del grupo de prueba, lo que indica una clasificación adecuada de la información por parte de los modelos, sin evidencia de sobreajuste (*overfitting*). Estos hallazgos subrayan la eficacia de los modelos de XGBoost en la imputación de datos en el contexto del análisis del mercado laboral.

Tabla 5. *Accuracy* modelos de imputación a través de XGBoost

XGBoost	
Nivel de instrucción	Experiencia
82,41%	83,03%

Al concluir la implementación metodológica de este trabajo, se logró consolidar una base de datos estructurada, robusta y coherente. Esta base incluye diversas variables clave para el análisis, cuyo detalle se expone en la Tabla 6.

Tabla 6. Estructura de la base de datos

Variable	Tipo de variable	Descripción
Empresa	Alfanumérico	Nombre de la empresa
Cargo solicitado	Alfanumérico	Nombre del cargo solicitado
Portal de empleo	Categórica	Nombre del portal de empleo
Fecha de inicio de publicación	Fecha	Fecha inicio de la publicación
Fecha de fin de publicación	Fecha	Fecha de fin de la publicación
Fecha de descarga	Fecha	Fecha de descarga de información
Sexo	Categórica	Sexo requerido en el anuncio de empleo
Tipo de contrato	Categórica	Tipo de contrato ofertado en el anuncio de empleo
Jornada laboral	Categórica	Jornada ofertada en el anuncio de empleo
Modalidad laboral	Categórica	Modalidad laboral ofertada en el anuncio de empleo
Salario	Categórica	Salario ofertado en el anuncio de empleo
Cargo ocupacional	Categórica	Descripción del cargo ocupacional según CIUO.08

[19] La correlación de Cramér es una medida estadística que cuantifica la asociación entre dos variables categóricas. Su valor varía de 0 a 1, donde 0 indica ausencia de asociación y 1 indica una asociación perfecta (Akoglu, 2018).
[20] Selección de forma aleatoria de los hiperparámetros óptimos para el adecuado desempeño del modelo de clasificación desde un conjunto de hiperparámetros establecido.
[21] k o fold contempla una división de los datos de entrenamiento en subconjuntos, en el que su valor indica el número de divisiones que serán realizadas sobre el conjunto de datos (Hastie et al., 2009).
[22] La métrica mide las predicciones correctas sobre el n° total de predicciones.

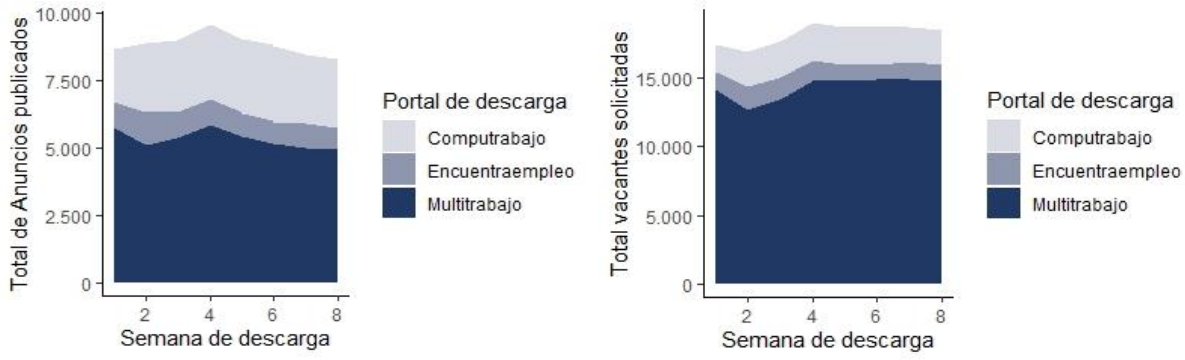
Rama de actividad	Categórica	Descripción de la rama de actividad de la empresa según CIIU.4.0
Código cargo ocupacional	Código numérico	Codificación del cargo ocupacional según CIUO.08 a 4 dígitos
Código rama de actividad	Código numérico	Codificación de rama de actividad según CIIU.4.0 a 4 dítigos
Nivel de instrucción	Categórica	Nivel de instrucción solicitado en el anuncio de empleo
Experiencia	Categórica	Experiencia mínima solicitada en el anuncio de empleo
Idioma	Categórica	Idioma requerido por el anuncio de empleo
Licencia de conducer	Categórica	Licencia de conducir solicitada por el anuncio de empleo
Discapacidad	Categórica	Personas con dificultades funcionales
Disponibilidad para viaje	Categórica	Disponibilidad para viajar
Provincia	Categórica	Provincia del cargo solicitado
Cantón	Categórica	Cantón del cargo solicitado
Código provincial	Código numérico	Codificación provincial DPA del cargo solicitado
Código canton	Código numérico	Codificación cantonal DPA del cargo solicitado
Total anuncios	Numérica	Número de anuncios descargados
Total vacantes	Numérica	Número de vacantes solicitadas en cada anuncio

3. Resultados

En este apartado se presentan diferentes estadísticas descriptivas alcanzadas a través de la construcción de una base de datos estructurada para análisis de anuncios de empleo y de vacantes laborales. Así, a través de la técnica de *web scraping* se pudo capturar un total de 70.700 anuncios de empleo y 194.814 vacantes abiertas para ser ocupadas. La descarga de información se realizó durante 8 semanas, las cuales se encuentran en el periodo del 18 de junio al 4 de septiembre del 2023.

Con la finalidad de analizar la consistencia de la información, en la Figura 3 se presenta el comportamiento de las descargas por portal de empleo y semana de descarga, donde se aprecia que durante el periodo analizado existe cierta estabilidad en el número de anuncios y vacantes disponibles en cada uno de los portales. Además, se pudo ver que mayoritariamente en promedio de las semanas de descarga, los registros provienen del portal de Multitrabajos (2023) con alrededor de 5.317 anuncios y 14.275 vacantes, seguidos de Computrabajo (2023) con aproximadamente 2.583 anuncios y 2.587 vacantes y, por último, el portal estatal Encuentraempleo (2023) muestra un total de 938 anuncios y 1.325 vacantes.

Figura 3. Descarga de anuncios y vacantes

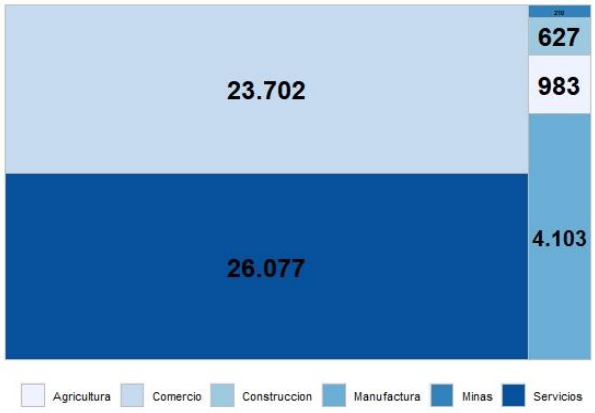


Este trabajo se planteó como meta final ofrecer una visión integral de la demanda laboral insatisfecha en el mercado laboral, analizando las vacantes disponibles en diversos anuncios de empleo. Para ello, se ha seleccionado un enfoque

centrado en los anuncios únicos de empleo identificados durante un periodo de ocho semanas. Esta selección se basa en variables relevantes como el nombre de la empresa, el cargo solicitado y la descripción de la oferta laboral; de este modo, se obtuvieron un total de 40.939 anuncios de empleo y 93.586 vacantes laborales. Mediante este enfoque, se buscó ofrecer un análisis detallado de las condiciones del mercado laboral ecuatoriano durante el periodo de las 8 semanas indagadas, examinando diversas dimensiones. Estas incluyen las ramas de actividad que más demandan trabajadores, los perfiles ocupacionales más buscados, los rangos salariales ofrecidos, así como los niveles de instrucción y experiencia requeridos por los empleadores. Este análisis pretende no solo identificar las tendencias actuales en el mercado laboral, sino también proporcionar una comprensión más profunda de las necesidades y expectativas de los empleadores en términos de calificaciones y competencias laborales.

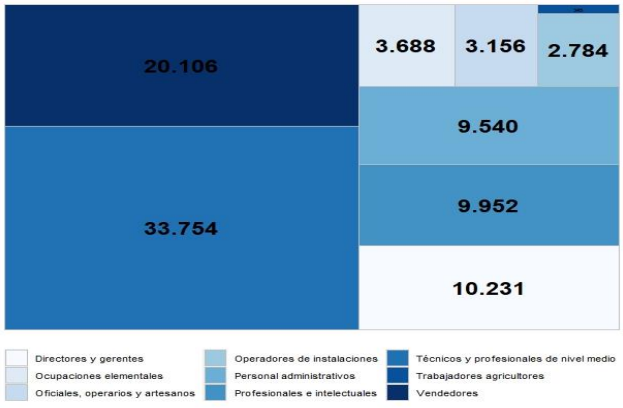
Las Figuras 4 y 5 ilustran de manera detallada el número de vacantes disponibles, clasificadas por rama de actividad y cargo ocupacional, respectivamente. Es notable que las empresas pertenecientes a sectores como servicios, comercio y manufactura, lideran la oferta de empleo en los portales analizados. De manera específica, el sector de servicios presenta 26.077 vacantes (27,8%), seguido por el comercio con 23.702 (25,3%) vacantes, y la manufactura con 4.103 (4,3%) vacantes.

Figura 4. Vacantes disponibles por rama de actividad de acuerdo al CIIU.4.0



En cuanto a los cargos ocupacionales, los más demandados por estas empresas en los portales de empleo son los técnicos de nivel medio con un total de 33.754 (36,1%) vacantes disponibles. A estos les siguen los vendedores, con 20.106 (21,4%) vacantes, y finalmente los directores y gerentes, con una oferta de 10.231 (10,9%) vacantes.

Figura 5. Vacantes disponibles por cargo ocupacional de acuerdo al CIUO.08



La Tabla 7 presenta una matriz de empleo de las vacantes laborales, relacionando las ramas de actividad con los cargos ocupacionales, y resaltando los roles más demandados en cada sector. En agricultura, predominan los operadores de

instalaciones con 240 vacantes (24.4%) y los técnicos de nivel medio con 211 vacantes (21.5%). En el sector del comercio, se observa una alta demanda de técnicos de nivel medio con 11.696 vacantes (49.3%), seguidos por vendedores con 3.859 vacantes (16.2%) y personal administrativo con 2.488 vacantes (10.5%). En construcción, aunque con un volumen menor de vacantes en total, sobresalen los técnicos de nivel medio con 179 vacantes (28.5%). La industria manufacturera muestra una inclinación hacia técnicos de nivel medio con 1.387 vacantes (33.8%) y profesionales e intelectuales con un total de 560 (13.7%). En el sector servicios, los roles más demandados son los técnicos de nivel medio con 7.879 vacantes (30.2%), seguidos por profesionales e intelectuales con 4.675 vacantes (17.9%) y personal administrativo con 3.236 vacantes (12.4%). Por su parte, en la rama de minas y canteras, a pesar de contar con un número limitado de vacantes, los directores y gerentes, técnicos de nivel medio y, profesionales e intelectuales son los más buscados, representando alrededor de 50 vacantes cada uno (apróx.25% cada uno).

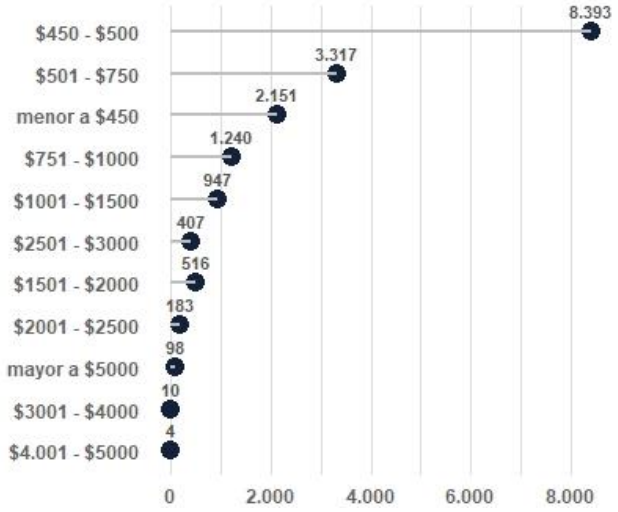
Tabla 7. Matriz de empleo

	Rama de actividad agregada de acuerdo al CIIU04								
Cargo ocupacional de acuerdo al CIUO08 a un dígito	Categoría de ocupación vs. Rama de actividad	Agricultura	Comercio	Construccion	Manufactura	Servicios	Minas	No clasificado*	Total
	Militares	-	6	-	-	16	-	3	25
	Directores	52	2.033	57	465	3.532	55	4.037	10.231
	Profesionales	91	1.613	102	560	4.675	51	2.860	9.952
	Técnicos	211	11.696	179	1.387	7.879	54	12.348	33.754
	Administrativos	114	2.488	33	327	3.236	11	3.331	9.540
	Vendedores	69	3.859	59	473	3.117	14	12.515	20.106
	Oficiales	148	774	94	339	971	8	822	3.156
	Operadores	240	358	78	266	906	14	922	2.784
	Agricultores	9	51	5	8	67	-	205	345
	Elementales	49	822	17	278	1.678	3	841	3.688
	No clasificado*	-	2	3	-	-	-	-	5
	Total	983	23.702	627	4.103	26.077	210	37.884	-

Nota:
*Corresponde a los registros de los cuales no fue posible clasificar la información de rama de actividad o cargo ocupacional a través de modelos de similitud de texto.

Además del análisis anterior, otro aspecto clave en la oferta de empleo es el salario. Sin embargo, es importante destacar que 76.320 (81,6 %) de las vacantes publicadas no especifican un sueldo concreto, dejándolo a convenir con el postulante. En este contexto, la Figura 6 ofrece una visión general de los salarios, basada en la información disponible en los anuncios de empleo. Para facilitar el análisis, se han establecido rangos salariales, tomando como base el Salario Básico Unificado (SBU) de Ecuador para 2023, que es de USD 450 y la cifra superior de USD 5.000. Los resultados muestran que 10.544 (11,2%) de las vacantes se sitúa en torno al SBU (menores o iguales al rango de USD 450 – USD 500), mientras que 4.557 (4,9%) vacantes se encuentran en el rango de USD 500 – USD 1.000.

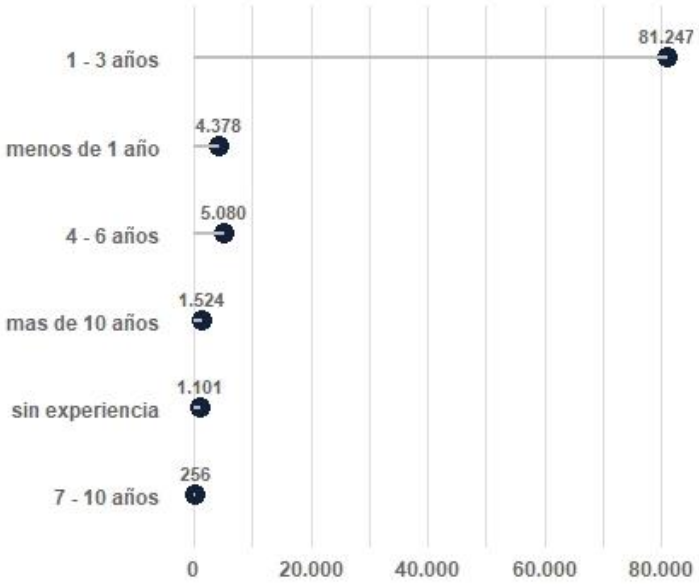
Figura 6. Vacantes disponibles por rangos salariales



Nota:
Para mejorar la visualización se ha eliminado a la categoría "a convenir".

En cuanto a la experiencia solicitada y el nivel de instrucción requerido, la Figura 7 revela que solo 1.101 (1,2%) vacantes no exigen experiencia previa, y 81.247 (86,8%) piden entre 1 y 3 años de experiencia.

Figura 7. Vacantes disponibles por experiencia solicitada



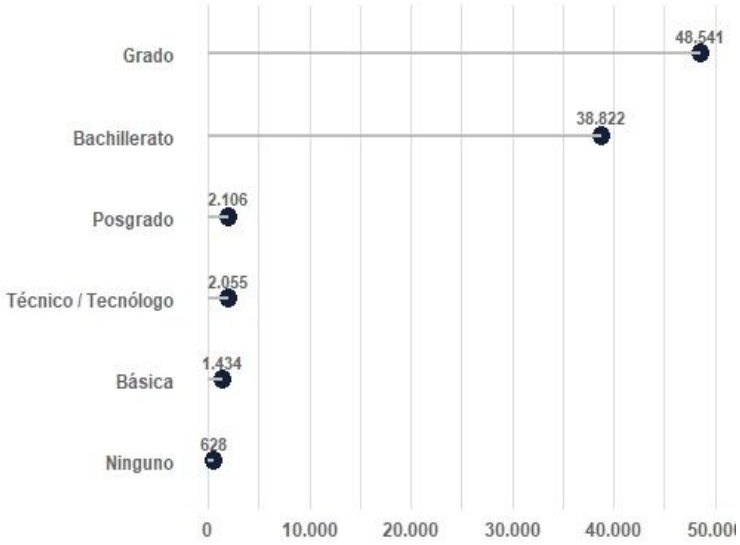
Al relacionar los salarios con la experiencia requerida, la Tabla 8 indica que la mayoría de las vacantes solicita entre 1 y 6 años de experiencia, con salarios que oscilan principalmente entre USD 450 y USD 750.

Tabla 8. Salario vs. Experiencia

	Experiencia (años)							
	Experiencia vs. Sueldo	Menos de 1 año	1 – 3	4 – 6	7 – 10	Más de 10 años	Sin experiencia	Total
Rangos salariales (USD)	menor a 450	226	1.811	31	-	83	-	2.151
	450 – 500	890	5.713	341	67	580	802	8.393
	501 – 750	189	2.468	362	6	137	155	3.317
	751 – 1.000	74	903	203	2	42	16	1.240
	1.001 – 1.500	33	660	238	8	4	4	947
	1.501 – 2.000	6	269	226	15	-	-	516
	2.001 – 2.500	-	115	33	34	1	-	183
	2.501 – 3.000	-	140	262	-	5	-	407
	3.001 – 4.000	-	3	5	2	-	-	10
	4.001 – 5.000	-	2	2	-	-	-	4
	Mayor a 5.000	40	56	-	-	2	-	98
	A convenir	2.920	69.107	3.377	122	670	124	76.320
	Total	4.378	81.247	5.080	256	1.524	1.101	-

La Figura 8 examina las vacantes según el nivel de instrucción demandado. Se observa que 48.541 (51,9%) vacantes requieren como mínimo un título de grado, y 38.822 (41,4%) solicitan un nivel mínimo de instrucción de bachillerato.

Figura 8. Vacantes disponibles por nivel de instrucción



Al vincular el salario con el nivel de instrucción requerido, la Tabla 9 muestra una concentración de vacantes en el rango salarial de USD 450 a USD 750, donde comúnmente se solicita un nivel de instrucción de bachillerato o de grado.

Tabla 9. Salario vs. Nivel de instrucción

	Nivel de instrucción							
	Salario vs. Nivel de instrucción	Ninguno	Basica	Bachillerato	Técnico - Tecnólogo	Grado	Posgrado	Total
Rangos salariales (USD)	Menor a 450	-	31	418	71	1.618	13	2.151
	450 – 500	488	347	4.386	756	2.336	80	8.393
	501 – 750	86	98	1.092	463	1.539	39	3.317
	751 – 1.000	35	23	238	104	793	47	1.240
	1.001 – 1.500	-	6	57	58	692	134	947
	1.501 – 2.000	-	-	21	2	400	93	516
	2.001 – 2.500	-	-	-	-	56	127	183
	2.501 – 3.000	-	-	-	2	20	385	407
	3.001 – 4.000	-	-	3	-	3	4	10
	4.001 – 5.000	-	-	-	-	-	4	4
	Mayor a 5.000	-	-	76	-	22	-	98
	A convenir	19	929	32.531	599	41.062	1.180	76.320
	Total	628	1.434	38.822	2.055	48.541	2.106	-

4. Conclusiones

Este trabajo se propuso desarrollar una metodología para crear una herramienta de software capaz de generar estadísticas sobre las vacantes en el mercado laboral, con el objetivo de generar cifras de demanda laboral y disminuir la brecha de información existente entre empleadores y trabajadores en Ecuador.

Como parte de la depuración se emplearon técnicas de *web scraping* y *web crawling* para extraer datos de los principales portales de empleo del país, como Computrabajo (2023), Multitrabajos (2023) y Encuentraempleo (2023), consolidando esta información en una base de datos estructurada. Además, se llevó a cabo un proceso intensivo de depuración y homologación de datos mediante técnicas de minería de texto y algoritmos basados en modelos de similitud de texto (*all-mpnet-base-v2 y Sentence-T5-Large*), aplicando clasificaciones estandarizadas como la DPA nacional (INEC, 2022), CIIU.4.0 (INEC, 2012a) y CIUO.08 (INEC, 2012b). Por último, se empleó el algoritmo de *machine learning* XGBoost para la imputación de variables de experiencia y nivel de instrucción.

Este enfoque mejoró significativamente el estudio previo realizado por Benítez et al. (2016), extendiendo la aplicación de *web scraping* a múltiples portales y utilizando métodos más avanzados para el procesamiento de datos. Además, se pudo generar variables relevantes para el análisis del mercado laboral, como son la rama de actividad y el cargo ocupacional hasta cuatro dígitos de desagregación del CIIU.4.0 y CIUO.08 respectivamente. Ofreciendo de esta manera información relevante para el análisis de la demanda laboral nacional.

La recolección de datos para este trabajo se llevó a cabo entre el 18 de junio y el 4 de septiembre de 2023, abarcando los tres portales de empleo. Se logró capturar un total de 70.700 anuncios, que representaron 194.814 vacantes. Tras un riguroso proceso de depuración y filtrado, se identificaron 40.939 anuncios únicos, equivalente a 93.586 vacantes. El análisis de estos datos reveló una concentración de oportunidades laborales en los sectores de servicios, comercio y manufactura, con una demanda predominante de cargos técnicos y profesionales de nivel medio, vendedores, y directores y gerentes. La mayoría de estas vacantes estaban dirigidas a candidatos con educación de nivel bachillerato o universitario, y con experiencia laboral de entre 1 y 3 años.

Al comparar estos hallazgos con los resultados obtenidos por Benítez et al. (2016), se constata que los requisitos laborales han permanecido relativamente constantes a lo largo del tiempo, particularmente en lo que respecta a los salarios ofertados, niveles de instrucción y experiencia requerida. Resulta notable que una proporción significativa de las ofertas laborales se continúe centrando en cargos de vendedores y asesores comerciales.

En conclusión, la herramienta desarrollada en este trabajo demuestra las ventajas de usar información de portales de empleo en línea para obtener datos relevantes del mercado laboral, a menudo en tiempo real y con costos operativos reducidos. A pesar de ciertas limitaciones de la técnica de *web scraping*, como la posible no representatividad de las cifras para un universo de análisis específico, su aplicación es esencial para generar estadísticas actualizadas y detalladas. Estos datos son cruciales para la creación de políticas públicas eficientes, cerrando la brecha informativa entre empleadores y trabajadores, y proporcionando cifras útiles sobre la demanda laboral, que son el eje central de este trabajo.

De cara al futuro, se planea expandir el alcance del software para abarcar una mayor cantidad de portales de empleo en línea, lo que permitirá mejorar la comprensión sobre las necesidades del mercado laboral nacional. Se considera también enriquecer el software con la adición de nuevas variables que ofrezcan una mirada más profunda a las condiciones y requisitos del mercado laboral, incluyendo características de las empresas y habilidades específicas solicitadas para diversos puestos. Sería beneficioso que el software evolucionara hacia una operación estadística regular, suministrando datos periódicos sobre las vacantes disponibles. Esto no solo ayudaría a la actualización de programas y currículos académicos, sino que también orientaría la formación profesional, alineándose de manera eficaz con las demandas del mercado laboral.

Bibliografía

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 91- 93.

Australian Government. (2023). *Jobs and Skills Australia*. Obtenido de https://www.jobsandskills.gov.au/work/internet-vacancy-index/methodology

Autor, D. (2015). Why Are There Still So Many Jobs? The History and Future of Workplace Automation. *Journal of economic perspectives, 29*(3), 3-30.

Baig, M., Shuib, L., y Yadegaridehkordi, E. (2019). Big Data Tools: Advantages and Disadvantages. *Journal of Soft Computing and Decision Journal of Soft Computing and Decision*, 14- 20.

Barcaroli, G., Fusco, D., Giordano, P., Greco, M., Moretti, V., Righi, P., y Scarno, M. (2016). ISTAT Farm Register: Data Collection by Using Web Scraping for Agritourism Farms. *ICAS VII Seventh International Conference on Agricultural Statistics*, 1- 8.

Benítez, D., Lucero, S., y Pazminño, A. (2016). *Elaboración de estadísticas de vacantes publicadas en internet. Una experiencia en Ecuador.*

Bergstra, J., Bengio, y Yoshua. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research 13* , 281-305.

Borjas, G. (2016). *Labor Economics.* New York: McGraw-Hill Education.

Cárdenas, Guataquí, y Montaña. (2015). *Metodología para el análisis de demanda laboral mediante datos de Internet: el caso colombiano.*

Cárdenas, J. (2020). *A Web-Based Approach to Measure Skill Mismatches and Skills Profiles for a Developing Country: The Case of Colombia.* Bogotá: Editorial Universidad del Rosario. doi:https://doi.org/10.12804/urosario9789587845457

Carrillo, P., y Vásconez, V. (2019). Caracterización de la demanda laboral de las empresas con información administrativa. *X-pendientes Económicos*, 1- 24.

Colombo, E., Mercorio, F., y Mezzanzanica, M. (2018). Applying machine learning tools on web vacancies for labour market and skill analysis. 1-30.

Computrabajo. (2023). Obtenido de https://ec.computrabajo.com/

Devlin, J., Chang, M.-W., Lee, K., y Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv (Cornell University)*. Obtenido de https://arxiv.org/pdf/1810.04805v2

Encuentraempleo. (2023). Obtenido de https://encuentraempleo.trabajo.gob.ec/socioEmpleo-war/paginas/index.jsf

Frid-Nielsen, S. (2019). Find my next job: labor market recommendations using administrative big data. *Association for Computing Machinery*, 408-412.

Geigle, G., Pfeiffer, J., Reimers, N., Vulić, I., y Gurevych, I. (2022). Retrieve Fast, Rerank Smart: Cooperative and Joint Approaches for Improved Cross-Modal Retrieval. *Transactions of the Association for Computational Linguistics, 10*, 503-521. doi:10.1162/tacl_a_00473

George, G., y Haas, M. (2014). Big Data and Management. *Academy of Management Journal, 57*(2), 321-326. doi:10.5465/amj.2014.4002

Gontero, S., y Menéndez, E. (2021). Macrodatos (Big Data) y mercado laboral. Identificación de habilidades a través de vacantes de empleo en línea. *CEPAL*, 1-54.

Hastie, T., Tibshirani, R., y Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction.* New York: Springer.

Hastie, T., Tibshirani, R., y Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. *New York: Springer*.

Henderson, M., Al-Rfou, R., Strope, B., László, L., Guo, R., Kumar, S., . . . Kurzweil, R. (2017). Efficient natural language response suggestion for smart reply. *arXiv (Cornell University)*. Obtenido de https://arxiv.org/pdf/1705.00652.pdf

Hermans, A., Beyer, L., y Leibe, B. (2017). In defense of the triplet loss for person Re-Identification. *arXiv (Cornell University)*. Obtenido de http://export.arxiv.org/pdf/1703.07737

Hirschberg, J., y Manning, C. D. (2015). Advances in natural language processing. *Science, 349*(6245), 261-266. doi:10.1126/science.aaa8685.

INEC. (2023). *Imputación de Información en Encuestas de Hogares a través de técnicas de Machine Learning: Análisis del caso ecuatoriano.*

Instituto Nacional de Estadística y Censos (INEC). (2012a). *Clasificación de Actividades Económicas (CIIU rev 4.0).*

Instituto Nacional de Estadística y Censos (INEC). (2012b). *Clasificación Nacional de Ocupaciones (CIUOO8).*

Instituto Nacional de Estadística y Censos (INEC). (2022). *CLASIFICADOR GEOGRÁFICO ESTADÍSTICO 2022.*

Kässi, O., y Lehdonvirta, V. (2018). Online labour index: Measuring the online gig economy for policy and research. *Technological Forecasting and Social Change, 137*, 241-248.

Khurana, D., Koli, A., Khatter, K., y Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications, 82*(3), 3713–3744. doi:10.1007/s11042-022-13428-4

Marchi, V., Apicerni, V., y Marasco, A. (2021). Assessing Online Sustainability Communication of Italian Cultural Destinations – A Web Content Mining Approach. *Information and Communication Technologies in Tourism* , 58- 69.

Maurer, S., y Liu, Y. (2007). Developing Effective E-Recruiting Websites: Insights for Managers from Marketers. *Business Horizons 50*, 305- 314.

Min, B., Ross, H., Sulem, E., Veyseh, A., Nguyen, T., Sainz, O., . . . Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: a survey. *ACM Computing Surveys, 56*(2), 1-40. doi:10.1145/3605943

Muennighoff, N., Tazi, N., Magne, L., y Reimers, N. (2022). MTEB: Massive Text Embedding Benchmark. *arXiv (Cornell University)*. doi:10.48550/arxiv.2210.07316

Multitrabajos. (2023). Obtenido de https://www.multitrabajos.com/empresas?utm_source=google&utm_medium=cpc&utm_campaign=B2B-GS-Brand&gclid=Cj0KCQiA2eKtBhDcARIsAEGTG40dkmkTdK90d6zs1YKV6Pe6OTKvxhgGTQBjw-JKYRmjOXeEltCX1ScaAi4fEALw_wcB

Nigam, H., y Biswas, P. (2021). From Web Scraping to Web Crawling. En A. Choudhary, A. Prakash Agrawal, y R. Logeswaran, *Applications of Artificial Intelligence and Machine Learning* (págs. 97- 112). Singapore: Springer Nature Singapore Pte Ltd.

Oancea, B., y Necula, M. (2019). Web scraping techniques for price statistics – the Romanian experience. *Statistical Journal of the IAOS 35* , 657- 667.

OECD. (2022). *Skills for the Digital Transition: Assessing Recent Trends Using Big Data.* Paris: PECD Publishing. Obtenido de https://www.oecd-ilibrary.org/sites/7d99dfbe-en/index.html?itemId=/content/component/7d99dfbe-en

Polidoro, F., Giannini, R., Lo Conte, R., y Rosetti, F. (2015). Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation. *Statistical Journal of the IAOS 31*, 165–176.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., . . . Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research, 21*(140), 1-67. Obtenido de https://jmlr.org/papers/volume21/20-074/20-074.pdf

Reimers, N., y Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. *arXiv (Cornell University)*. Obtenido de https://arxiv.org/pdf/1908.10084.pdf

Rosati, G. (2021). Métodos de Machine Learning como alternativa para la imputación de datos perdidos: Un ejercicio en base a la Encuesta Permanente de Hogares. *Estudios del Trabajo*.

Rubin, D. (1976). Inference and Missing Data. *Biometrika Trust, 63*(3), 581-592.

Sozzi, A. (2018). Measuring Sustainability Reporting using Web Scraping and Natural Language Processing. *NTTS*.

ten Bosch, O., Windmeijer, D., van Delden, A., y van den Heuvel, G. (2018). *Web scraping meets survey design: combining forces.*

Tlamelo, Maupong, Mpoeleng, Semong, Mphago, y Tabona. (2021). A survey on missing data in machine learning. *Journal of big data*, 1-37.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention Is All You Need. *Advances in neural information processing systems, 30*.

www.ecuadorencifras.gob.ec