Modelado laboral de los egresados de la Facultad de Ciencias Informáticas de la Universidad Técnica de Manabí (Ecuador)

Juan C. Cevallos, Martha C. Escobar, Jessenea E. Falcones y Wilson J. Cevallos

Universidad Técnica de Manabí. Av. Urbina y Che Guevara. Portoviejo – Ecuador

(correo-e: ing.cevallos@gmail.com; marthitae65@gmail.com; truvifalcon@hotmail.com; wilson.cevallos@gmail.com)

Recibido Abr. 7, 2021; Aceptado Jun. 12, 2021; Versión final Ago. 10, 2021, Publicado Dic. 2021

Resumen

El objetivo de este estudio fue desarrollar un modelo de minería de datos para identificar patrones de empleabilidad de egresados, que contribuya a mejorar la tasa de inserción laboral afín a la formación profesional. Este estudio responde a la preocupación por la empleabilidad de profesionales. Se utilizó la metodología KDD (Knowledge Discovery in Databases) por etapas definidas y simplificadas, algoritmos y técnicas de árboles de decisión, redes neuronales y regresión para determinar la interacción de los egresados en un entorno laboral. Se identificó el algoritmo red neuronal que se ajustaba para resolver la problemática por exactitud, precisión, sensibilidad, tiempo de ejecución, rendimiento y facilidad para interpretación de datos y resultados del modelo generado. Se concluye que las variables predictoras significantes son: 1) uso de conocimientos y habilidades en su trabajo actual, 2) el año de graduación y si buscó trabajo al término de su carrera, y 3) labora actualmente (variable dependiente).

Palabras clave: modelado laboral; empleabilidad; egresados; aprendizaje supervisado; minería de datos; árboles de decisión

Modeling of the labor market for graduates of the Faculty of Computer Science at the Technical University of Manabi (Ecuador)

Abstract

The main objective of the present study was to develop a data mining model to identify employability patterns of graduates. The model aims to improve job placement rate related to vocational training. This study is in response to concerns about employability of professionals. The KDD (Knowledge Discovery in Databases) methodology was used with defined and simplified stages, algorithms, decision tree techniques, neural networks, and regressions to determine the interaction of graduates in the work environment. The neural network algorithm was adjusted for accuracy, precision, sensitivity, execution time, performance, and data and results interpretation for the generated model. In conclusion, the significant predictor variables were: 1) use of knowledge and skills in the current job, 2) the graduation year and if graduates searched for jobs after graduation, and 3) currently employed (dependent variable).

Keywords: job model; employability; graduates; supervised learning; data mining; decision tree

INTRODUCCIÓN

Las Instituciones de Educación Superior (IES) deben atender los requerimientos de las empresas y la sociedad, que permiten abrir espacio para la inserción laboral de sus egresados. De igual forma, articular la oferta docente, investigación y vinculación con la sociedad, a la demanda académica, necesidades de desarrollo local, regional y nacional, innovación y diversificación de profesiones y grados académicos, tendencias del mercado ocupacional local, regional y nacional. La Universidad Técnica de Manabí (UTM) se preocupa por vincular a los graduados con el mercado y ofertas de trabajo, debe crear enlaces entre las empresas públicas y privadas; sin embargo, este propósito no se logra totalmente.

Las IES diseñan carreras para formar profesionales competentes que respondan y den soluciones a los problemas de la sociedad; asumen el compromiso y responsabilidad social de ofrecer mecanismos que contribuyan a encontrar fuentes de trabajo idóneas para los graduados. Sin embargo, existen brechas entre las habilidades adquiridas en la universidad y las habilidades que exigen los empleadores (Cranmer 2006). Sudakova et al., (2018), consideran que la construcción de un modelo de optimización para el proceso de formación, permite estimar y calcular un escenario de la producción de la universidad teniendo en cuenta la demanda del mercado laboral. Este modelo tuvo como principal resultado la reducción del número total de desempleados egresados de la universidad y la mejora en la calidad de los profesionales en formación. Agregó que se debe tener en cuenta la mano de obra demandada por el mercado para incrementar las especializaciones en la formación universitaria. En este contexto, la UTM, cuenta con el sistema informático de Seguimiento a Graduados e Inserción Laboral, para recabar información pertinente de los graduados. Sin embargo, no se contaba con los conocimientos, técnicas y tecnologías que permitieran explorar los datos de manera automática o semiautomática. Por ello, se concibe la necesidad de encontrar patrones interesantes, hasta ahora desconocidos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto.

Por esta razón, el objetivo de este estudio fue, desarrollar un modelo de minería de datos para identificar patrones de empleabilidad de egresados e incrementar la tasa de inserción laboral afín a la formación profesional. Cada vez más estudios están investigando técnicas de minería de datos para predecir la empleabilidad (Razón et al., 2019; Aviso et al., 2021; Romero y Ventura, 2010; Peña, 2014). Denila (2020) analizó técnicas de minería de datos para evaluar el empleo y fortalecer las decisiones universitarias. Sudakova et al., (2018) creó un modelo para optimizar la estructura de la formación y salida de los graduados de las universidades para atender las necesidades del mercado laboral. Tomlinson, (2012) indagó acerca de las formas en que los estudiantes y graduados construyen su empleabilidad y comienzan a gestionar la transición de la universidad al trabajo. Rahman et al., (2017) propuso un modelo que aporta a la gestión universitaria a ejecutar planes a largo plazo para formar graduados capaces que satisfagan las necesidades de la industria. Adekitan y Salau, (2019) analizaron el impacto de la ingeniería en el desempeño de estudiantes después de su graduación. Fontalvo-Herrera et al (2018) analizaron la aplicación de minería de datos en la calificación de planes de estudio de programas universitarios.

Desde el sustento teórico, las variables de estudio se obtienen de datos cuantitativos o cualitativos, sus relaciones producen el conocimiento inducido o patrones de comportamiento identificados, convirtiéndose en un elemento vital de análisis predictivo. Una gran cantidad de datos, está emparentado con lo que se ha conocido como *Big Data*, (Crawford et al., 2015; Hernández et al., 2017). En este proceso computacional Gullo (2015), opina que una vez que se dispone de los datos, se induce conocimiento, se emplean técnicas de aprendizaje automático y análisis predictivos.

El análisis predictivo es un área de la minería de datos, consiste en la extracción de información existente en datos válidos. Sirve para mejorar rendimiento en datos ruidosos o desequilibrados (Dietterich, 2000). Se aplica para predecir tendencias y búsqueda de patrones de comportamiento e inducir conocimiento, puede aplicarse sobre cualquier evento desconocido, en el pasado, presente o futuro (Espino, 2017). Se basa en el reconocimiento de vínculos entre variables en acontecimientos pasados y predecir posibles resultados futuros. Permite estimar eventos a raíz del análisis de datos históricos y presentes a través de métodos estadísticos y de reconocimiento de patrones (Marz, 2015). Un patrón es una entidad a la que se le puede dar un nombre y que está representada por un conjunto de propiedades medidas y las relaciones entre ellas (Watanabe, 1985). En el caso de predecir patrones de empleabilidad de egresados, si se cruzan datos de perfiles de los profesionales, requerimiento laboral de las empresas, salarios, experiencia, título obtenido, edad; se puede inferir en esos patrones y modelos predictivos.

El modelo predictivo implica el uso de algoritmos, como red neuronal, regresión logística, regresión lineal, entre otros. Tiene como objetivo clasificar los datos en una de varias clases categóricas predefinidas o utilizar campos seleccionados de datos históricos para predecir campos objetivo (Hirji, 2002). Se utiliza para predecir probabilidades de que un egresado pueda encontrar un empleo afín a la formación recibida. Una vez introducidos los datos del egresado y se aplique el modelo predictivo se obtendrá una calificación que indicará

la probabilidad de que se produzca la situación estudiada por el modelo. En un modelo predictivo, el grado no es siempre al ciento por ciento; esto se debe a que, aunque se haya repetido un patrón de comportamiento en el pasado, no tiene por qué repetirse. Sin embargo, siempre será mejor predecir con ayuda del modelo, que simplemente adivinar (Espino, 2017).

De igual manera, los modelos descriptivos permiten encontrar y definir relaciones interesantes en los datos utilizando aprendizaje no supervisado (García, 2018). A diferencia de los modelos predictivos que se centran en predecir el comportamiento de un egresado en particular, los modelos descriptivos identifican diferentes relaciones entre los egresados y las empresas, asignándoles una categoría según su preferencia en el mercado laboral o su aspiración salarial. El sistema de reconocimiento de patrones identifica el patrón como miembro de una clase definida (clasificación supervisada); y, asigna el patrón a una clase todavía no definida (clasificación no supervisada, agrupamiento).

El aprendizaje supervisado permite deducir una función a partir de los datos de entrenamiento, cuyo objetivo es predecir el valor de un atributo (etiqueta), de un conjunto de datos (atributos) (Moreno et al., 2005). Se desarrolla en dos fases: 1) Entrenamiento: Construcción de un modelo usando un subconjunto de datos con etiquetas conocidas; 2) Prueba: Prueba del modelo sobre el resto de los datos. El aprendizaje no supervisado, es exploratorio y describe la estructura de los datos para su análisis. Para aplicar estos algoritmos de minería de datos se requiere la construcción de actividades previas dedicadas a preparar los datos de entrada. El proceso consiste en: 1) Determinación de objetivos; 2) Preparación de datos; 3) Transformación de datos; 4) Minería de datos; 5) Análisis de resultados; 6) Asimilación de conocimiento (Cabena et al., 1997). Las principales técnicas (supervisadas y no supervisadas) de minería de datos, son: árboles de decisión y redes neuronales artificiales.

El árbol de decisión está formado por un conjunto de nodos de decisión (interiores) y de nodos-respuesta (hojas) que representan las etiquetas de la clase y los valores de los atributos de esta instancia se comparan con las ramas del árbol de decisión que conducen a esas etiquetas de clase. El objetivo es crear un modelo que predice el valor de una variable de destino en función de diversas variables de entrada; estos valores pueden ser nominales o con atributos de valores numéricos, mediante la aplicación de los algoritmos adecuados (ID3 y C4.5) (Berdón et al., 2021). Las redes neuronales artificiales, son sistemas que aprenden y se forman a sí mismos, en lugar de ser programados de forma explícita; y, sobresalen en áreas donde la detección de soluciones o característica es difícil de expresar con la programación convencional. En este contexto, se validó el rendimiento del modelo usando los datos de aprendizaje y el resultado del algoritmo; y se identificó la línea de tendencia que determina el pronóstico de inserción laboral de los futuros egresados de la Facultad de Ciencias Informáticas de la Universidad Técnica de Manabí.

METODOLOGÍA

Para el desarrollo del modelo se realizó una revisión bibliográfica. En el proceso se usó la metodología de minería de datos, por las etapas que posee, claramente definidas y simplificadas, beneficios directos que residen en el valor de la información y el conocimiento que se extraiga de datos almacenados en las fuentes. Se explotó la metodología KDD, esta selección se basó en el análisis comparativo de las metodologías más utilizadas y que no estén aplicadas en proyectos similares a esta investigación. Las metodologías CRISP-DM, SEMMA y KDD, aportan por la forma particular de llevar el proceso de extraer patrones en forma de reglas o funciones a partir de los datos, para que el usuario los estudie. En la actualidad, el término de KDD y minería de datos se utilizan como referencia al proceso completo de descubrimiento de conocimiento, fue propuesta por Fayyad et al., en 1996 y proponen 6 fases: análisis previo, selección, pre-procesamiento/limpieza, transformación/reducción, minería de datos e interpretación/evaluación de datos.

Se realizó el análisis previo o inicial de aproximación al problema. Continuó la fase de selección, después de haber identificado el conocimiento relevante, prioritario y definidas las metas del proceso KDD, se creó un conjunto de datos objetivo, seleccionando una muestra representativa de este, sobre el cual se realizó el proceso de descubrimiento. Esta fase terminó con el entregable "Datos a Analizar". En la siguiente fase de pre-procesamiento/limpieza, se analizó la calidad de los datos, se aplicaron operaciones básicas como la remoción de datos ruidosos, se seleccionaron estrategias para el manejo de datos desconocidos, datos nulos, datos duplicados y técnicas estadísticas para su reemplazo. Esta fase terminó con el entregable "Datos Preprocesados". En la fase de transformación/reducción, se buscaron características útiles para representar los datos según la meta del proceso. Se utilizaron métodos de reducción de dimensiones o de transformación para reducir el número efectivo de variables bajo consideración (Fayyad et al., 1996). Esta fase terminó con el entregable "Datos Transformados".

La fase de minería de datos tuvo como objetivo la búsqueda y descubrimiento de patrones insospechados y de interés, aplicando tareas de descubrimiento como clasificación (Quinlan, 1986), patrones secuenciales y asociaciones, entre otras; terminó con el entregable "Patrones". En la fase de interpretación/evaluación de

datos, se interpretaron los patrones descubiertos. Esta etapa incluyó la visualización de los patrones extraídos, la remoción de los patrones redundantes o irrelevantes y la traducción de los patrones útiles en términos que fueran comprensibles para el usuario. Se consolidó el conocimiento descubierto para incorporarlo, en posteriores acciones, documentarlo y reportarlo a las partes interesadas; así mismo, para verificar y resolver conflictos potenciales con el conocimiento previamente descubierto. Se finalizó el procedimiento con el entregable “Conocimiento”.

Por último, la etapa de selección de herramientas pretendió tener datos íntegros, completos y consistentes, para ello fue necesario el uso de herramientas para explorar, explotar la información, analizar los datos para la detección de errores, transformación de datos y eliminación de duplicados. Se evaluaron las herramientas para la limpieza y transformación de los datos, ofreciendo como resultado: exactitud, integridad, entereza, validez, consistencia, uniformidad, densidad y unicidad. Después de la comparación de varias herramientas, se seleccionó la herramienta *OpenRefine*.

RESULTADOS Y DISCUSIÓN

En este apartado se describen los procesos de diseño y ejecución del modelo y los resultados obtenidos en torno a las seis fases descritas en la metodología. Los resultados de la contribución se convirtieron en una herramienta importante para analizar el camino de los egresados, su incorporación a instituciones públicas o privadas. De igual forma, la situación real de la igualdad de género en el trabajo, importancia de la formación académica para acceder progresiva y rápidamente al entorno laboral y mejoramiento de la empleabilidad de graduados en el contexto de la experiencia. Todas orientadas al principio de calidad que contribuiría al cumplimiento del criterio de pertinencia y vinculación con la sociedad requerido por el Consejo de Aseguramiento de la Calidad y de la Educación Superior (CACES), que considera a la pertinencia como el principio que se refiere a las capacidades que tiene una carrea para responder y articularse a las demandas del entorno.

Fuentes de datos (análisis previo): En esta fase se accedió a los datos de los graduados, base de entrenamiento y prueba del modelo. La base estaba diseñada en *PostgreSQL 11.5*, es un sistema avanzado de gestión de bases de datos relacionales de objetos (ORDBMS), derivado del sistema *Berkeley Postgres*. La identificación de las fuentes de datos lleva cierta cantidad de tiempo, cuando no se posee diccionario de datos o el Modelo Entidad Relación (MER); además, comprender cuáles son las categorías de datos y las fuentes de origen de estos, ayuda a establecer información relevante al objetivo. Se analizaron las siguientes bases de datos que se describen en la tabla 1.

Tabla 1: Base de datos y tablas de la base

Bases de datos	Tablas de la base	
Base de datos db_sga	esq_datos_personales.personal	contiene los registros de los egresados de la Facultad de Ciencias Informáticas de la Universidad Técnica de Manabí
Base de datos utmssg	Título	contiene los registros de los títulos que otorga la UTM
	Preguntas	almacena el detalle de las preguntas que se realizaron a los egresados
	pregunta_opcion	almacena el detalle de las respuestas, relacionada con las preguntas

Selección: Con el planteamiento de objetivos, se precisó la recolección de datos, tipo de extracción, atributos de entrada y salida, para obtener un subconjunto o datos de calidad, que mejore la eficiencia del proceso de minería de datos. Se seleccionaron las variables relevantes en los datos, buscando su simetría, normalidad y correlación existente en la información. Para recuperar la información dentro de la base de datos, se trabajó con el lenguaje de consulta o SQL, proporcionados por la Escuela de Tecnologías de la Información, de la Facultad de Ciencias Informáticas, de la Universidad Técnica de Manabí, datos almacenados en el Sistema de Seguimiento de Graduados e Inserción Laboral, SSGIL-UTM, que registra los datos completos del graduado y profesional de todas las cohortes de la carrera, el período de obtención de la información fue junio 2005 a diciembre 2017; exportado en un subconjunto de datos en formato.xlsx. La data contó con un total de 929 graduados, 22 958 registros, de las variables objetivo (aquellas que se quiere predecir, calcular o inferir); y, de las variables independientes (sirven para hacer el cálculo o proceso). Estos registros contienen información relevante de: género, fecha de graduación; si se encuentra laborando actualmente; si labora en una empresa pública o privada; modalidad con la que se graduó; si tiene postgrado; actividad profesional que realiza; si está relacionada con su campo de estudio; cuando consiguió su primer trabajo; rango de ingresos mensuales en dólares; capacitación profesional en los últimos 12 meses, entre otros campos.

Teniendo en cuenta el estado de la técnica, los costes de aplicación, la naturaleza, el alcance, el contexto y los fines del tratamiento, así como los riesgos de probabilidad y gravedad variables para los derechos y las libertades de las personas físicas, se establecen las medidas técnicas y organizativas apropiadas para la mitigación de riesgos y garantizar el nivel de seguridad adecuado al riesgo existente, que incluye la seudonimización y el cifrado de datos personales en referencia a todo aquello que contiene: Información de identificación directa como la cédula, el nombre y el apellido.

Pre-procesamiento/limpieza: Una vez que se extrajo la información y se exportó la data seleccionada; se continuó con la limpieza de ruido o datos anómalos y tratamiento de datos faltantes, para garantizar la utilidad de estos. La herramienta *OpenRefine*, permitió cumplir con el objetivo de las cinco tareas esenciales en la limpieza de los datos: Eliminar registros duplicados. Anonimizar datos (identificar y ocultar la información sensible). Separar varios valores contenidos en el mismo campo. Analizar la distribución de valores a lo largo de un conjunto de datos. Agrupar diferentes representaciones de la misma realidad. Como primer paso se ejecutó el programa *openrefine.exe*, se abrió una página web en el navegador de Internet a través de la dirección <http://127.0.0.1:3333/>, e importamos el archivo DATOS_GRADUADOS.xlsx para empezar a explorar y modificar. En el apartado "*Facet/Filter*" se filtran los campos que no inciden en el desarrollo de la presente investigación y se remueve.

Transformación/reducción: Una vez culminado el proceso de Pre-procesamiento/limpieza, continuó la mejora en la calidad de los datos con transformaciones que involucran reducción de dimensiones (disminuir la cantidad de variables del conjunto de datos) o bien transformaciones. Se transformó la fecha de graduación para obtener el año en que se graduó el egresado, la data final tuvo un total de 632 graduados e igual número de registros.

Minería de datos: Fue la fase más importante de la metodología KDD, se construyó el modelo predictivo, de clasificación o segmentación aplicando los algoritmos y técnicas seleccionadas al conjunto de datos en búsqueda de los patrones de utilidad. Los modelos predictivos pretendieron estimar valores futuros o desconocidos de variables de interés, que se denominan variables objetivo, dependientes o clases, usando otras variables denominadas independientes o predictivas. Para el desarrollo de esta metodología se seleccionaron los siguientes atributos relevantes para la clase que labora actualmente: CBET: Carrera como base al empezar a trabajar. BTTC: Búsqueda de trabajo al término su carrera. UCHTA: Uso de conocimientos y habilidades en su trabajo actual. IM: Ingresos mensuales. PT: Primer trabajo. IL: Institución que Labora. APF: Actividad Profesional afín a la formación. IF: Instrucción formal. MG: Modalidad de graduación. Año_graduacion: Año de Graduación.

En esa fase se empleó la herramienta *Weka* 3.8.3 para desarrollar el modelo; es un software abierto y consiste en una colección de algoritmos de aprendizaje autónomo en Java para tareas de minería de datos que permite el pre-procesado, clasificación, regresión, agrupación, asociación y definición de reglas o visualizaciones, entre otras. Se utilizó el enfoque predictivo para diseñar modelos que se aplican en datos futuros, esencialmente para predecir comportamientos. En inteligencia artificial se indican como modelos de aprendizaje supervisado. Las variables pueden ser categóricas y numéricas. Para la creación del modelo predictivo se utilizaron unidades de muestra disponibles con atributos y comportamiento conocido, a este conjunto de datos se le denominó conjunto de entrenamiento. Por otro lado, se utilizó una serie de unidades de otra muestra con atributos similares, de las cuales no se conocía su comportamiento, a este conjunto de datos se le denominó conjunto de prueba. Para ejecutar los métodos de clasificación en *Weka* (redes neuronales, regresión y árboles de decisión), se aplicaron los siguientes pasos: preprocesado, clasificación (algoritmo básico de aprendizaje J48, algoritmo de redes neuronales, algoritmo Regresión lineal).

Preprocesado: Se definió el origen de los datos. Aunque el formato por defecto de *Weka* es el *arff*, se puede seleccionar el archivo CSV (Archivos separados por comas o tabuladores) que exportamos en la fase anterior mediante la herramienta *OpenRefine*. El menú desplegable permitió variar el atributo de referencia, se representó en color para contrastar ambos atributos. *Visualize all* mostró todas las gráficas pertenecientes a todos los atributos. En la segunda pestaña (zona superior) del explorador entramos en el modo clasificación. Los algoritmos de clasificación tienen como fin clasificar una variable dentro de las categorías de una clase, aplicando la premisa todas las clases detectadas son disjuntas o diferentes entre sí. Una variable solo puede pertenecer a una clase. Los datos cargados se clasificaron por varios métodos y se analizaron los siguientes algoritmos: J48, C4.5, Red Neuronal y Regresión Lineal.

J48: En *Weka* es el algoritmo básico de aprendizaje de árboles de decisión C4.5. Es una secuencia de decisiones que se organizan de manera jerárquica, como ramas de un árbol. Se empleó para tareas de clasificación, agrupación y pronóstico. Si se predicen categorías se denominan árboles de clasificación; si son numéricos y se pretende predecir, se llaman árboles de regresión. Este algoritmo construyó el árbol de decisión de manera descendente, con la interrogante, ¿qué atributo es el que debería ser colocado en la raíz del árbol? Cada atributo fue evaluado con un test estadístico. Se seleccionó la raíz, se agregó una rama para cada valor del atributo correspondiente; y, el conjunto de entrenamiento se ordenó en los nodos apropiados.

En la clasificación de la nueva instancia, los atributos especificados por los nodos son evaluados; inició su recorrido por el nodo raíz y en forma descendente se recorrieron las ramas del árbol que son los valores de los atributos en la instancia dada. El algoritmo se interrumpió cuando los ejemplos de entrenamiento compartieron el mismo valor para el atributo que se ensayaba o la cobertura mínima del número de ejemplos por cada nodo está por debajo de cierto umbral, en este punto la etiqueta asociada a la hoja se asigna a la nueva instancia como su categoría. El algoritmo C4.5, mide qué tan bien un atributo dado separa el conjunto de entrenamiento conforme a las clases. El funcionamiento del algoritmo en pseudocódigo, se puede resumir en: 1) Comprobar los casos base. 2) Para cada atributo a : Encontrar la ganancia de información normalizada de la división de a . 3) Dejar que a_{best} sea el atributo con la ganancia de información normalizada más alta. 4) Crear un nodo de decisión que divida a_{best} . 5) Repetir sublistas obtenidas por división de a_{best} , y agregar estos nodos como hijos de nodo.

Para ejecutar este algoritmo, dentro del área del clasificador del árbol de decisión, se seleccionó la opción *trees – J48* para clasificar el modelo. Se empezó por la configuración del modo de entrenamiento; se escogió la opción *Use training set* (en este caso se usa para hacer el test el mismo conjunto que el de entrenamiento); Weka entrenó el método con todos los datos disponibles y luego los aplicó otra vez sobre los mismos. El menú desplegable permite seleccionar un atributo de la muestra. Este atributo es el que actúa como resultado real de la clasificación (*Labora Actualmente*), habitualmente este atributo suele ser el último.

El factor de confianza para la poda, influyó notoriamente en el tamaño y capacidad de predicción del árbol, para cada operación de poda, definió la probabilidad de error que se permite a la hipótesis de que el empeoramiento debido a esta operación es significativo. Cuanto más baja sea esa probabilidad, se exigirá que la diferencia en los errores de predicción antes y después de podar sea más significativa para no podar. El valor por defecto de este factor fue 25%, y conforme va bajando se lograron más operaciones de poda para llegar a árboles cada vez más pequeños. Se construyó el árbol de decisión con los parámetros por defecto del algoritmo J48. El modelo clasificó las instancias de forma correcta en 97.7848% que representa a un total de 618 instancias de las 632 en total, el número de hojas: 27 y el tamaño del árbol: 34. Su estructura se visualizó en forma de gráfico. J48 de la caja *Result-list*, evidencia el árbol de decisión podado, que se muestra en la Figura 1.

Se observa que después de la poda se incrementa la tasa de error del árbol calculada sobre los ejemplos de entrenamiento del 97.7848% que ha clasificado las instancias de forma correcta, a un 93.038% y del 2.2152% aumenta a un 6.963% las instancias clasificadas incorrectamente. Entre los aspectos más relevantes tenemos: El modelo ha clasificado las instancias de forma correcta en 93.038% que representa a un total de 588 instancias de las 632 en total. La matriz de confusión, con a y b representan las etiquetas de clase. Aquí hay 632 instancias, por lo que los porcentajes y los números se suman, $aa + ab = 159 + 35 = 194$ como la clase NO y $bb + ba = 429 + 9 = 438$ como la clase SI. El porcentaje de instancias correctamente clasificadas a menudo se llama exactitud o precisión de la muestra, de las cuales se puede aseverar que de la clase NO, 159 se clasificaron correctamente, mientras 35 fueron incorrectamente; por otro lado, la clase SI, 429 se clasificaron de forma correcta, mientras que 9 fueron incorrectamente. El número de hojas: 9 y el tamaño del árbol: 14. Para facilitar la interpretación de resultados en esta investigación, se trabajó con el árbol podado.

MultilayerPerceptron: Es el nombre que se le da en *Weka* al algoritmo de redes neuronales artificiales, el cual realizó tareas de clasificar, predecir y agrupar, tiene la desventaja de trabajar con datos numéricos; sin embargo, las variables categóricas se suelen discretizar para aplicar estos algoritmos. Este algoritmo se aplicó para discretizar un conjunto de valores numéricos en rangos de datos. Como parámetros tomó los índices de los atributos discretizar y el número de particiones en que pretende dividir los datos. Otra de las configuraciones importantes en el filtro Discretizar, se presentan en la tabla 2.

Entre los aspectos relevantes el modelo ha clasificado las instancias de forma correcta en 96.2025% que representa 608 instancias de las 632 en total. La matriz de confusión, con a y b representan las etiquetas de clase. Aquí hay 632 instancias, por lo que los porcentajes y los números se suman, $aa + ab = 182 + 12 = 194$ como la clase NO y $bb + ba = 426 + 12 = 438$ como la clase SI. El porcentaje de instancias correctamente clasificadas a menudo se llama exactitud o precisión de la muestra, de las cuales se puede aseverar que de la clase NO, 182 se clasificaron correctamente, mientras que 12 incorrectamente. Por otro lado, tenemos la clase SI, 426 se clasificaron de forma correcta, mientras que 12 incorrectamente.

Regresión Lineal, es el nombre que *Weka* da al algoritmo de Regresión, el cual permitió analizar la relación existente entre la variable dependiente o de respuesta y un conjunto de variables independientes o predictoras; se expresa como una ecuación que predice la variable de respuesta como una función lineal de los parámetros. El objetivo de la regresión fue seleccionar los parámetros del modelo que minimizan la suma de los errores al cuadrado y se representa mediante una ecuación lineal $Y = a * X + b$, donde: Y – Dependent Variable. a – Slope. X – Independent Variable. b – Intercept.

Fig. 1: Árbol podado

Tabla 2: Atributos y características del filtro Discretizar

Atributos	Características
Debug	si se establece en verdadero, el filtro puede generar información adicional a la consola.
doNotCheckCapabilities:	si se cambia a verdadero, las capacidades del filtro no se verifican antes de crearse.
attributeIndices:	especifica el rango de atributos para actuar. Esta es una lista de índices de atributos separados por comas, con valores válidos "first" y "last". Especifique un rango inclusivo con "-".
useBinNumbers:	al cambiar a true permite usar números de bin en lugar de rangos para atributos discretos.
invertSelection:	establece el modo de selección de atributos. Si es falso, solo los atributos (numéricos) seleccionados en el rango serán discretizados; si es verdadero, solo los atributos no seleccionados serán discretizados.

Para implementar este algoritmo fue necesario cambiar el origen de los datos de la estructura inicial; se dejaron dos atributos: anio_graduación como variable independiente y labora_actualmente como variable dependiente. La data contenía un total de 13 registros, desde el año 2005 hasta el 2017. La regresión se calculó con la siguiente fórmula: $\% \text{ Labora Actualmente} = -4.9112 * \text{Anio_Graduación} + 9954.765$. El coeficiente de relación fue 0.9133%. Este coeficiente, permitió entender qué tanta correlación existe entre el Anio_Graduación y si labora actualmente. En este contexto tenemos: Correlación perfecta: Cuando el resultado de coeficiente es igual a 1 o -1. Correlación fuerte: Cuando el resultado es mayor a 0.5 y menor que 1 (correlación positiva) o menor a -0.5 y mayor que -1 (correlación negativa). Correlación débil: Valores que están entre -0.5 y 0.5.

El resultado indica que se tiene una correlación fuerte. Puede existir un margen de error de un 32.5878 %, debido a que todas las medidas están afectadas en algún grado por un error experimental según las imperfecciones inevitables del instrumento de medida o las limitaciones impuestas por nuestros sentidos que deben de registrar la información. Con la fórmula anterior se efectuó la regresión, para ello se utilizó Microsoft Excel para evaluar, se muestra la proyección de los próximos 13 años, en la tabla 3.

Tabla 3: Proyección de los próximos 13 años para inserción laboral

Anio_Graduación	% Labora Actualmente	
2005	100	
2006	100	
2007	100	
2008	88.88	
2009	89.58	
2010	89.74	
2011	91.89	
2012	89.58	
2013	52.27	
2014	63.25	
2015	59.46	
2016	50	
2017	45	
2018	43.96	$= -4.9112 * 2018 + 9954.765$
2019	39.05	$= -4.9112 * 2019 + 9954.765$
2020	34.14	$= -4.9112 * 2020 + 9954.765$
2021	29.23	$= -4.9112 * 2021 + 9954.765$
2022	24.32	$= -4.9112 * 2022 + 9954.765$
2023	19.41	$= -4.9112 * 2023 + 9954.765$
2024	14.50	$= -4.9112 * 2024 + 9954.765$
2025	9.58	$= -4.9112 * 2025 + 9954.765$
2026	4.67	$= -4.9112 * 2026 + 9954.765$
2027	-0.24	$= -4.9112 * 2027 + 9954.765$
2028	-5.15	$= -4.9112 * 2028 + 9954.765$
2029	-10.06	$= -4.9112 * 2029 + 9954.765$
2030	-14.97	$= -4.9112 * 2030 + 9954.765$

Interpretación/evaluación de datos: Se analizaron los resultados de los patrones obtenidos en la fase de Minería de datos, mediante técnicas de visualización y de representación, con el fin de generar conocimiento que aporte mayor valor a los datos. En esta fase se evaluaron los resultados con los responsables del CSGIL – UTM. La validez de contenido se aseguró por la selección sistemática de los métodos estadísticos y configuración por defecto de cada algoritmo. Para todos los algoritmos aplicados se estableció como método de la validación y entrenamiento completo; esta opción evaluó el clasificador del mismo conjunto sobre el que se construyó el modelo predictivo para determinar el error, que en este caso se denomina "error de resustitución", los resultados obtenidos se muestran en la tabla 4. La matriz de confusión del algoritmo Redes Neuronales evidencia concentración de las instancias en la diagonal lo que indica que en el modelo predominan los verdaderos positivos y los verdaderos negativos; un mayor porcentaje de las instancias son clasificadas correctamente. Sin embargo, este algoritmo es el que se toma más tiempo para realizar el entrenamiento.

Los valores sensibilidad favorecen a la red neuronal, que obtuvo 0.938% para la clase NO y 0.973% para la clase SI. La sensibilidad mide que tan completos son los resultados e indica la capacidad del estimador para discriminar los casos positivos, de los negativos. En el algoritmo C4.5, el árbol de decisión se convierte en reglas de clasificación; es decir, se crea una regla por cada camino de la raíz a las hojas; y, cada una representa una conjunción lógica (operador AND), mientras que el árbol completo es una disyunción de conjunciones (operador OR), en este contexto tenemos como resultado las siguientes reglas en la tabla 5.

Tabla 4. Resultados de la ejecución de los algoritmos C4.5, Redes Neuronales y Regression Lineal

Modelo	Instancias clasificadas correctamente	Instancias clasificadas incorrectamente	Tiempo necesario para probar el modelo	Clase	Tasa de verdaderos positivos	Tasa de falsos positivos	Precisión	Sensibilidad	Matriz de confusión	
									A	B
Algoritmo C4.5	588	44	0.00 s	NO	0.820	0.021	0.946	0.82	159	35
				SI	0.979	0.180	0.925	0.97	9	429
Algoritmo Redes Neuronales	608	24	0.03 s	NO	0.938	0.027	0.938	0.938	182	12
				SI	0.973	0.062	0.973	0.97	12	426
Algoritmo Regresión Lineal	N/A	N/A	0.01 s	N/A	N/A	N/A	N/A	N/A	N/A	N/A
				N/A	N/A	N/A	N/A	N/A	N/A	N/A

Tabla 5: Algoritmo C4.5: Operadores y Reglas de clasificación

Operador	Reglas de clasificación (Labora Actualmente)
AND	SI Uso de conocimientos y habilidades en su trabajo actual es NO APLICA AND Anio_Graduación <= 2012 AND Búsqueda de trabajo al termino su carrera es SI AND Anio_Graduación > 2010 AND Carrera como base al empezar a trabajar Es MUY POCO
OR	Uso de conocimientos y habilidades en su trabajo actual es MEDIANAMENTE OR Uso de conocimientos y habilidades en su trabajo actual es MUCHO OR Uso de conocimientos y habilidades en su trabajo actual es MUY POCO

La ventaja del uso de árboles de decisión permanece en la facilidad de comprensión de los resultados, creando un marco que implica la perspicacia del conocimiento, es de gran aceptación para resolver problemas con incertidumbre y falta de exactitud de datos, como se aprecia en la Figura 2. El algoritmo de Regresión Lineal ayudó a predecir la variable dependiente (y), con los valores de las variables independientes (X) permitió pronosticar el porcentaje de inserción laboral a los futuros egresados de la Facultad de Ciencias Informáticas de la Universidad Técnica de Manabí en los próximos trece años, se muestra en la Figura 3. En esta figura, validamos el rendimiento del modelo usando los datos de aprendizaje y el resultado del algoritmo; se puede dibujar una línea de regresión llamada línea de tendencia, cuya inclinación determina un comportamiento decreciente en el pronóstico de inserción laboral de los futuros egresados de la Facultad de Ciencias Informáticas de la Universidad Técnica de Manabí, existiendo una correlación fuerte entre las variables que componen el modelo.

Fig. 2: Análisis del árbol por niveles

Fig. 3: Pronóstico de inserción laboral a los futuros egresados de la Facultad de Ciencias Informáticas

El perfil de egreso de la Escuela de Tecnologías de la Información de la Facultad de Ciencias Informáticas, Universidad Técnica de Manabí, con énfasis en desarrollo de aplicaciones *Web* y móviles, administración y gestión de redes y comunicaciones, seguridad informática, con la utilización eficiente de recursos humanos, financieros y tecnológicos; de acuerdo con los resultados del algoritmo de Regresión Lineal, la demanda de estos perfiles y por ende la inserción laboral en los próximos años en empresas públicas y/o privadas registran un descenso acelerado.

Los resultados enfatizan que la atención debería concentrarse en una transformación del mercado laboral. El avance de nuevas tecnologías contribuye a que las empresas apuesten cada vez más por la innovación y busquen perfiles profesionales digitales; y, se crea la necesidad de nuevos perfiles, de esta manera se evita la sobrepoblación de profesionales. Por lo tanto, estos resultados permiten el análisis y creación de nuevos perfiles y especialidades de acuerdo con estudios de pertinencia, según los modelos de evaluación de carrera e institucional dispuestos por los organismos de aseguramiento de la calidad de la educación superior del país. Los patrones de empleabilidad encontrados en este estudio, le permiten a la Facultad de Ciencias Informáticas, plantear reformas en el diseño curricular, para dar respuesta a los requerimientos del entorno laboral. Actualizar contenidos teóricos y prácticos de acuerdo con el avance tecnológico de las empresas, así como estudios de empleabilidad y bolsas de empleo, para lograr una inserción laboral afín a la formación profesional que se oferta.

Del conjunto de métricas evaluadas, se evidencia que el modelo entrenado con redes neuronales logró un excelente desempeño en precisión y calidad de los resultados; por tanto, es el más efectivo. Autores como Sudakova et al., (2018), se han preocupado de analizar la inserción laboral de profesionales y han planteado modelos e identificado beneficios relacionados con la reducción del número total de desempleados egresados de la universidad y mejora de la calidad de los estudiantes en formación; en el desarrollo del modelo se sugirió realizar una optimización de escenarios de producción de la universidad teniendo en cuenta la mano de obra y demanda del mercado de graduados.

En el estudio de Barón (2010), se concluyó que los modelos presentados servían para identificar el futuro profesional del graduado en el corto plazo, en relación a la demanda y oferta de mano de obra calificada. Afirmó, que el salario no es la única medida del desenvolvimiento económico de los graduados; también lo es la probabilidad de encontrar un empleo de calidad en los mercados formales de trabajo. En este contexto, Ramanathan et al., (2020) creó el modelo de clasificación y demostró que el funcionamiento del algoritmo se basó en el principio de segregación de la fuerza laboral en función de las calificaciones y habilidades profesionales. De igual forma, Teijeiro et al., (2013) demostró que las universidades deben cambiar su enfoque tradicional y hacer un esfuerzo especial para ayudar a sus estudiantes a desarrollar aquellas competencias que mejor fomenten la empleabilidad.

CONCLUSIONES

De acuerdo a los resultados de este estudio y de su discusión, se pueden extraer las siguientes conclusiones: 1) se identificaron las fuentes de datos y variables de interés que contribuyeron al éxito y desarrollo de la metodología de minería de datos, con atributos independientes o predictores como CBET, BTTC, UCHTA, IM, PT, IL, APF, IF, MG para su clase Anio_graduacion o atributo dependiente, que aseguran la calidad de la información; 2) uno de los patrones repetitivos, tendencia o regla que presentan los egresados es que tienen una alta probabilidad de trabajar en cargos que no son afines a la profesión y no emplean los conocimientos adquiridos en su gran mayoría; no se realiza una evaluación de conocimientos y vocación profesional en el ingreso a la carrera; 3) durante el desarrollo de la metodología de minería de datos se implementaron diferentes algoritmos de clasificación, tales como J48, Redes Neuronales y Regresión Lineal; siendo el más efectivo para el modelo desarrollado Redes Neuronales por su precisión y calidad de los resultados. El algoritmo que mayor facilidad tiene para interpretar e interactuar con los resultados, es el J48 o también denominado árbol decisiones por su entorno gráfico en la salida de los resultados; 4) se implementó el algoritmo Regresión Lineal que permite pronosticar la inserción laboral a los futuros egresados de la Facultad de Ciencias Informáticas de la Universidad Técnica de Manabí, dando como resultado una tendencia cuya inclinación es decreciente en el presagio de años venideros. La apertura de otras especialidades en la carrera brindará otras probabilidades de inserción laboral.

REFERENCIAS

- Adekitan, A., y Salau, O., The impact of engineering students' performance in the first three years on their graduation result using educational data mining, <http://doi.org/10.1016/j.heliyon.2019.e01250>, Heliyon, 5(2), (2019)
- Aviso, K.B., Demetrio III, F.P., Janairo, J.I., y Lucas, R.I., What university attributes predict for graduate employability?, <https://doi.org/10.1016/j.clet.2021.100069>, Cleaner Engineering and Technology, 2, 1-8 (2021)
- Baron, J., Primeras experiencias laborales de los profesionales colombianos: probabilidad de empleo formal y salarios, *Lecturas de Economía* 76, 55-86 (2012)
- Berón, E., Mejía, D., y Castrillón, O., Principales causas de ausentismo laboral: una aplicación desde la minería de datos, <http://dx.doi.org/10.4067/S0718-07642021000200011>, Información Tecnológica, 32(2), 11-18 (2021)
- Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., y Zanasi, A., *Discovering data mining from concept to implementation*, Perentice Hall, (1998)
- Cranmer S., Enhancing graduate employability: best intentions and mixed outcomes, <https://doi.org/10.1080/03075070600572041>, Studies in Higher Education, 31(2), 169–184 (2006)
- Crawford, M., Khoshgoftaar, T., Prusa, J., Richter, A., y Najada, A., Survey of review spam detection using machine learning techniques, <http://doi.org/10.1186/s40537-015-0029-9>, Journal of big data, 2(23), 2-24 (2015)
- Denila, P., Delima, A., Vilchez, R., Analysis of IT graduates employment alignment using C4.5 and naïve bayes algorithm, <http://doi.org/10.30534/ijatcse/2020/106912020>, International Journal of Advanced Trends in Computer Science and Engineering, 9(1), 745–752 (2020)
- Dietterich, T.G., Ensemble methods in machine learning. In: multiple classifier systems, Springer, Berlin Heidelberg, 1–15, (2000)
- Espino, T.C., Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas open source que permiten su uso, Tesis, Universitat Oberta de Catalunya, (2017)
- Fayyad, U., Piatetsky-Shapiro, G., y Smyth, P., KDD process for extracting useful knowledge from volumes of data, *Communications of the ACM*, 39(11), 27-34 (1996)
- Fontalvo-Herrera, T., Delahoz, E., y Mendoza-Mendoza, A., Aplicación de minería de datos para la clasificación de programas universitarios de ingeniería industrial acreditados con alta calidad en Colombia, *Información Tecnológica*, 29(3), 89-96 (2018)
- García, A.M., Modelos descriptivos basados en aprendizaje supervisado para el tratamiento de grandes volúmenes de datos y flujos continuos de datos, XVIII Conferencia de la Asociación Española para la inteligencia artificial, 1402-1407 (2018)
- Hernández-Leal, E., Duque-Méndez, N., Moreno-Cadavid, J., Big data: una exploración de investigaciones, tecnologías y casos de aplicación, *Tecnológica*, 20(39) (2017)
- Gullo, F., From patterns in data to knowledge discovery: what data mining can do, *Physics Procedia* 62, 18-22 (2015)
- Hirji, K., Exploring data mining implementation, *Communications of the ACM*, 44(7), 87-93 (2001)
- Marz, J., *Principles and best practices of scalable real-time data systems*, Manning Publications Co (2015)

- Moreno, M., Quintales, L., García, M., y Martín J., Aplicación de técnicas de minería de datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software, I Congreso Español de Informática, CEDI-2015, 2-4 (2005)
- Peña-Ayala, A., Educational data mining: a survey and a data mining-based analysis of recent works, <https://dx.doi.org/10.1016/j.eswa.2013.08.042>, Expert Systems with Applications, 41(4), 1432-1462 (2014)
- Quinlan, J.R., Induction of decision trees, Machine Learning Journal 1, 81-106 (1986)
- Rahman, N.A., Tan, K.L., y Lim, C.K., Predictive analysis and data mining among the employment of fresh graduate students in HEI, <https://doi.org/10.1063/1.5005340>, AIP Conference Proceedings 1891,020007 (2017)
- Ramanathan, L., Sreenivas Dasagrandhi, S., Hrushikar, K., y Chorghe, R., Classification modelling and analysis of professionals skillsets, <https://doi.org/10.1016/j.matpr.2020.10.669>, Materials today: proceedings, (2020)
- Razón, J., Ortega, F., y Lozano, A., Impacto de la gestión de certificaciones internacionales en la inserción laboral de los ingenieros electromecánicos, Revista Electrónica Anfei Digital 5 (10), 1-8 (2019)
- Romero, C., y Ventura, S., Educational data mining: a review of the state of the art, <https://doi.org/10.1109/TSMCC.2010.2053532>, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601-618 (2010)
- Sudakova, A.E., Argakov, G.A., y Shorikov, A.F., Optimization of the graduates labour market: dynamic modeling, Russian and foreing experience, IFAC Papers OnLine, 51(32), 401-406 (2018)
- Teijeiro, M., Rungo, P., y Freire, M., Graduate competencies and employability: the impact of matching firms' needs and personal attainments, <https://doi.org/10.1016/j.econedurev.2013.01.003>, Economics of Education Review, 34, 286-295 (2013)
- Tomlinson, M., Graduate employability: a review of conceptual and empirical themes, Higher Education Policy, (25), 407-431 (2012)
- Watanabe, S., Pattern recognition: human and mechanical, New York, John Wiley y Sons (1985)