Inteligencia artificial y datos masivos en archivos digitales sonoros y audiovisuales

Perla Olivia Rodríguez Reséndiz
Coordinadora

Q335
I57

Inteligencia artificial y datos masivos en archivos digitales sonoros y audiovisuales / Coordinadora Perla Olivia Rodríguez Reséndiz. - México: UNAM. Instituto de Investigaciones Bibliotecológicas y de la Información, 2020.

xviii, 182 p. - (Tecnologías de la información)
ISBN:
Investigación realizada gracias al programa DGAPA - PAPIIT IT400118.

1. Inteligencia artificial - Procesamiento de datos. 2. Internet de las cosas. 3. Archivos sonoros. 4. Big data. I. Rodríguez Reséndiz, Perla Olivia, coordinadora. II. ser.

Diseño de portada: Oscar Fernando Arcos Casañas

Imágenes:
Envato Elements
(https://elements.envato.com/es-419/)

Primera edición, 2020
D.R. © UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
Ciudad Universitaria, 04510, México D.F.
Impreso y hecho en México
ISBN: En trámite

Publicación dictaminada

Propuesta metodológica y de análisis computacional para identificar el proceso fotográfico en fotografías históricas del siglo XIX y XX

GUSTAVO LOZANO SAN JUAN
Instituto de Investigaciones Estéticas. Universidad Nacional Autónoma de México.

RODRIGO COLÍN RIVERA
Laboratorio Audiovisual de Investigación Social, Instituto Mora.

OBJETIVO

El objetivo de este proyecto es diseñar una metodología de trabajo que permita a los archivistas, historiadores, conservadores y otros profesionales de los archivos, identificar correctamente la técnica de las fotografías producidas durante los siglos XIX y XX.

ANTECEDENTES

Como sabemos, el catálogo es el instrumento que describe ordenadamente y de forma individualizada las unidades documentales de una serie o conjunto documental (Heredia, 1991), las operaciones necesarias para la catalogación de una fotografía son: la elaboración de un lenguaje documental, el análisis de las unidades documentales y la sistematización de la información (Boadas, 2001). El análisis documental se refiere tanto al contenido como al continente, y el aspecto más relevante a documentar en este último caso, es la identificación de la técnica o proceso fotográfico.

Figura 1. Diversos procesos fotográficos

En la bibliografía sobre el tema, la identificación del proceso fotográfico no se presenta como un procedimiento sistemático compuesto por pasos claramente delimitados, sino como la aplicación de habilidades de observación y reconocimiento que son adquiridas de manera empírica a lo largo de años de práctica y estudio. Bajo el planteamiento

tradicional, es obligatorio que dichas habilidades se acompañen de un conocimiento sólido de la cronología de los procesos fotográficos y de sus principales hitos tecnológicos, requerimientos que plantean una curva de aprendizaje pronunciada.

En los cursos de capacitación y actualización existentes, la identificación de los procesos fotográficos se enseña de forma casi personalizada en grupos de tamaño pequeño y en clases de duración limitada. El instructor es generalmente un profesional de la conservación especializado en fotografías y de manera invariable se requiere de una colección de estudio variada que permita a los participantes desarrollar las habilidades de observación y análisis descritas previamente.

Tabla 1. Procesos fotográficos utilizados en los siglos XIX y XX

Imágenes de cámara	Impresiones	Negativos	Transparencias
Daguerrotipo	Cianotipo	Negativo de colodión húmedo	Transparencia plata gelatina sobre vidrio
Ambrotipo	Albúmina	Negativo de placa seca de gelatina	Transparencia procesos de pantalla aditiva
Ferrotipo	Colodión de impresión directa	Plata gelatina sobre nitrato de celulosa	Transparencia cromógena sobre acetato de celulosa
	Plata gelatina de impresión directa	Plata gelatina sobre acetato de celulosa	Transparencia cromógena sobre poliéster
	Platinotipo	Plata gelatina sobre poliéster	
	Impresión plata gelatina	Cromógeno sobre acetato de celulosa	

	Impresión por difusión de plata	Cromógeno sobre poliéster	
	Impresión cromógena		
	Impresión por difusión de colorantes		
	Impresión por blanqueo de colorantes		

JUSTIFICACIÓN

El proceso fotográfico se refiere tanto a los materiales como a la manera en que estos se combinan -artesanal o industrialmente- para crear la fotografía. Su documentación es importante porque da cuenta de la evolución tecnológica de prácticas extintas hoy en día y porque algunos procesos requieren de medidas de conservación especiales, como por ejemplo, el almacenamiento a bajas temperaturas para las impresiones, transparencias y negativos cromógenos o las medidas de prevención y combate de incendios para los negativos con soporte de nitrato de celulosa.

Desafortunadamente los vehículos tradicionales para la enseñanza–aprendizaje de los conocimientos y las habilidades requeridos para la identificación del proceso fotográfico, presentan varias limitantes que impiden su diseminación entre los profesionales de los archivos.

DESARROLLO

La propuesta que aquí se plantea pretende aliviar algunas de las limitaciones antes mencionadas, ya que se trata de una metodología de fácil acceso para los profesionales de los archivos, por medio de la cual es posible identificar el proceso de una fotografía de interés,

entre una gama de 30 alternativas utilizadas a lo largo de los siglos XIX y XX. Algunas de sus características más relevantes son que se pueden aplicar de forma individual sin necesidad de un instructor, no es necesaria, una colección de estudio, no depende del estudio previo de la historia técnica de la fotografía. Adicionalmente, permite a usuarios que aplican la metodología por primera vez, obtener resultados positivos y facilitar la diseminación de un conocimiento de acceso limitado.

A continuación se describe la manera en que esta propuesta fue desarrollada.

Investigación bibliográfica

En la primera etapa de este trabajo se hizo una búsqueda bibliográfica que permitiera identificar las principales características físicas que definen a cada uno de los procesos fotográficos (Reilly, 1986) (Sepiades, 2003) (Aasbø, 2003) (Barra, 2005) (Image Permanence Institute, 2019).

Tabla 2. Características físicas.

Soporte primario	Tonalidad
Iluminación	Brillo
Polaridad	Superficie
Tono	Texto
Fecha	Deterioro
Estratigrafía	Tonalidad
Magnificación	Particularidades del Objeto

Síntesis de la información

Posteriormente se sintetizó la información y se capturó en formato tabular, en las filas se colocaron los diferentes procesos fotográficos, en las columnas las diferentes características físicas a identificar y en las celdas en donde ambas se intersectan se colocaron las caracterís-

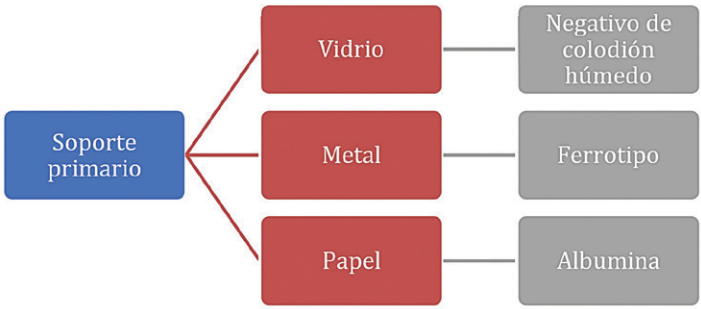
ticas concretas que definen a cada proceso. Por ejemplo, para el caso del proceso fotográfico negativo de colodión húmedo, la característica soporte primario corresponde a la opción vidrio, para el proceso ferrotipo corresponde a la opción metal y para el proceso albumina corresponde a la opción papel.

Tabla 3. Ejemplo de información en formato tabular.

	Soporte primario
Negativo de colodión húmedo	Vidrio
Ferrotipo	Metal
Albumina	Papel

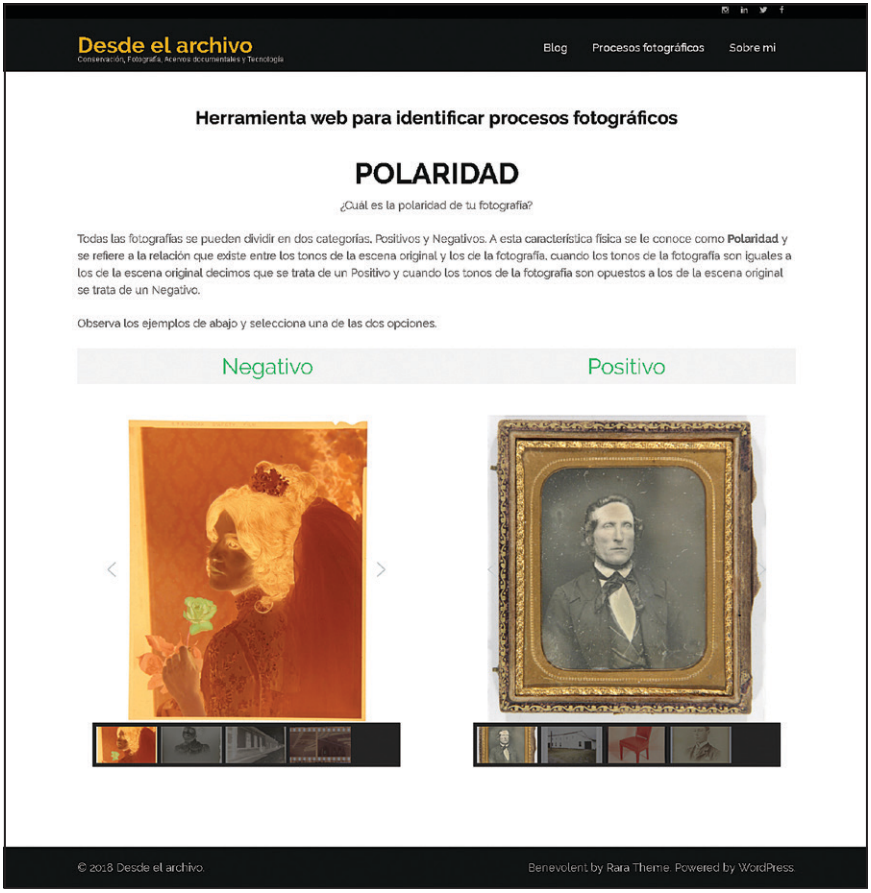
Esta misma información fue trasladada posteriormente a un árbol de decisiones, un modelo predictivo utilizado en las ciencias de datos para clasificar entidades con base en sus atributos; así el soporte primario que en la tabla se encontraba como encabezado de la segunda columna se convirtió en un nodo de decisión del cual se derivan tres ramas: vidrio, metal y papel, que corresponden a las opciones posibles para esa característica, finalmente los tres procesos fotográficos que en la tabla se encontraban en la primera columna se convierten en las hojas del árbol y corresponden a la clasificación final.

Figura 2. Ejemplo de información en formato de árbol de decisiones.



El modelo del árbol permitió elaborar un cuestionario (http://desde-elarchivo.com/id-procesos-fotograficos/) que a través de una serie de preguntas dirigidas guían al usuario paso a paso a través de la metodología hasta llegar al conjunto de fotografías que le permiten identificar el proceso fotográfico de su fotografía.

Figura 3. Interfaz de usuario de la página web.



En su estado actual el árbol cuenta con 40 nodos, 96 ramas y 30 hojas por lo que optimizarlo es sumamente importante. Para ello se propo-

ne también un análisis mediante conjuntos de datos representativos con diversas propiedades de las fotografías que serán procesados mediante herramientas y algoritmos de inteligencia artificial. El propósito es generar un árbol de decisiones alternativo a la metodología archivística para encontrar similitudes y determinar las propiedades de mayor relevancia al clasificar las fotografías.

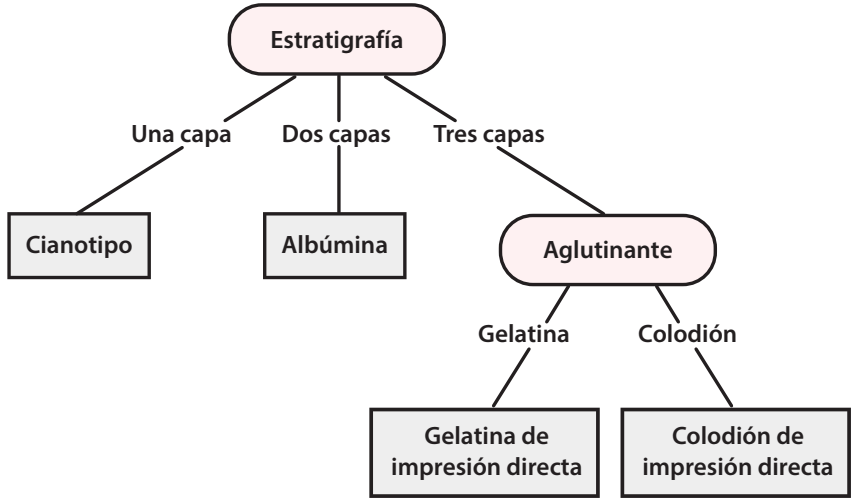
El tipo de análisis computacional propuesto deriva de una rama de la inteligencia artificial denominada Machine Learning (ML) o aprendizaje de máquina, donde el énfasis es aprender de los datos, sin que esta capacidad de aprendizaje esté explícitamente programada. Y de esta manera, utilizar los árboles de decisión como un modelo de predicción de un valor objetivo basado en un conjunto de datos variables.

La generación de árboles de decisión de manera automática o asistida por algoritmos computacionales permite desarrollar tareas de clasificación, cuya aplicación es realizar un análisis computacional para la identificación del proceso fotográfico en fotografías históricas del siglo XIX y XX. Al referirnos al término *identificación*, y bajo este contexto, es válido indicar que el término *predicción* es empleado como sinónimo, ya que el propósito es aproximar (con el menor margen de error posible) un conjunto de datos a un modelo de árbol de decisión que nos permita realizar dicha identificación fotográfica y además, brinde un punto de comparación en cuanto a las características físicas más relevantes de este conjunto particular de fotografías.

Hacer predicciones

Al revisar otro ejemplo simplificado de árbol de decisión para la identificación de procesos fotográficos, se aprecia la raíz en la parte superior y el árbol crece hasta los nodos terminales u hojas en la parte inferior. Esta representación es equivalente a tener el nodo raíz del lado izquierdo y hacerlo crecer hacia la derecha.

Figura 4. Ejemplo sencillo de árbol de decisión con nodo raíz en la parte superior.



Es interesante notar que cada uno de los nodos intermedios (en color rosa) del árbol de decisión representa una característica física de la fotografía mientras que las hojas (en color gris) representan procesos fotográficos que determinan de manera definitiva la categoría a la que pertenece. Como se puede apreciar, utilizar esta estructura de datos para tareas de clasificación es bastante claro e intuitivo, en contraste con otros modelos como las redes neuronales, considerados modelos de *caja negra* ya que usualmente es difícil de explicar en términos simples por qué y cómo son hechas las predicciones o clasificaciones (Gerón, 2017).

Ventajas y limitaciones

Como se ha mostrado, una ventaja importante de los árboles de decisión es que el proceso de clasificación es lo suficiente fácil de llevar a cabo teniendo la estructura del árbol, y se puede realizar manualmente sin la necesidad de una computadora. Los datos de entrada que se procesan usualmente derivan de conjuntos o bases de datos que requieren poca intervención y se pueden exportar en formatos de archi-

vo convenientes como CSV (valores separados por coma). Otra ventaja importante es que el desempeño es bueno con cantidades razonables de poder de cómputo; si se tiene un conjunto de datos grande, el proceso de aprendizaje y construcción del modelo es manejable (Bell, 2015).

A pesar de las ventajas también es importante identificar inconvenientes, y uno de los más importantes es el hecho de que se pueden crear modelos muy complejos generados a partir de conjuntos de datos complejos. La manera de contrarrestar este efecto es revisar y acotar los valores a usar en categorías, lo cual producirá un modelo más refinado y concreto. Hay que estar consciente de que este proceso no siempre es viable, de ser el caso, el modelo resultante puede ser mucho más grande de lo esperado. Puede ser que utilizar otro método de inteligencia artificial se adapte mejor a las necesidades del conjunto de datos, por ejemplo, redes neuronales o máquinas de soporte vectorial.

Algoritmos

ID3

El algoritmo ID3 (Iterative Dichotomiser 3) fue inventado por Ross Quinlan, investigador en ciencias de la computación en el área de minería de datos y teoría de la decisión, para crear árboles a partir de bases de datos. Al calcular la entropía --unidad de medida del desorden en un conjunto de datos, según la teoría de la información-- para cada atributo del conjunto de datos, esto permite dividir en subconjuntos basados en el valor de mínima entropía, y construir de manera recursiva cada nodo del árbol de decisión.

En conjunción, ID3 utiliza también el método de ganancia de información, la medida de las diferencias de entropía antes y después de que un atributo divide en subconjuntos de datos, para determinar el nodo raíz en cada llamada recursiva del algoritmo.

C4.5

Ross Quinlan efectuó mejoras en su algoritmo original y creó el algoritmo C4.5 que de igual manera emplea el método de ganancia de

información, pero hace más evidente que el árbol resultante puede ser usado como modelo de clasificación.

Entre las mejoras más notables respecto a ID3 se encuentra la habilidad de trabajar con atributos continuos, como valores numéricos o fechas. También permite trabajar con valores de atributos vacíos o "huecos" en la base de datos sin que esto afecte la ganancia de información de un atributo durante la construcción del árbol. Y por último, el árbol creado con C4.5 es recortado después de su creación, es decir, algunas ramificaciones del árbol son modificadas por nodos terminales cuando contribuyen a simplificar el árbol (Bell, 2015).

Proceso de generación automática de árboles de decisiones

Es usual en el área de ML, que el ciclo de acciones a tomar para generar modelos a partir de conjuntos de datos involucra los siguientes pasos: obtención de datos, preparación de éstos, ejecución de la heurística o algoritmo y presentar los resultados.

Se emplea el software de código abierto WEKA (https://www.cs.waikato.ac.nz/ml/weka) para la generación de árboles de decisión. Dicho software cuenta con una colección de algoritmos y herramientas de pre procesado de datos que permiten trabajar de manera flexible con conjuntos de datos, facilita la experimentación, la obtención de resultados estadísticos y visualización tanto de los datos de entrada como de los resultados obtenidos.

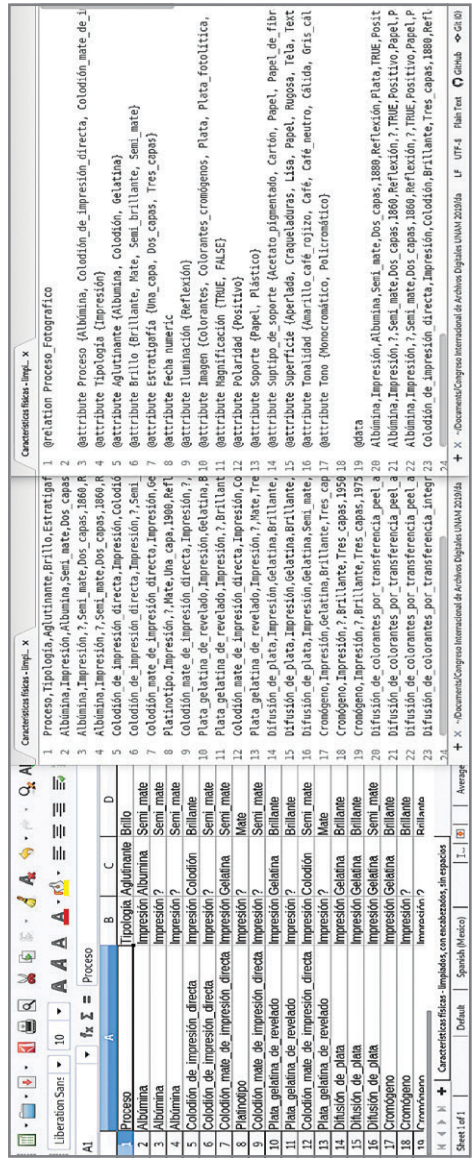
Obtención de datos

Los datos que se emplearon en este trabajo son una acotada selección de diversas características fotográficas de la colección de estudio perteneciente al área de conservación del Archivo Fotográfico Manuel Toussaint del Instituto de Investigaciones Estéticas de la UNAM. Todas las características físicas de los procesos fotográficos a identificar se registraron en una hoja de cálculo que sirve de insumo inicial en esta etapa del proceso.

Figura 5. Muestra de la hoja de cálculo que da cuenta de las propiedades y atributos fotográficos.

Clasificaciones		Atributos comunes				Atributos particulares							
Proceso	Tipología	Soporte primario	Iluminación	Polaridad	Tono	Fecha	Estratigrafía	Magnificación	Tonalidad	Brillo	Superficie	Particularidades Objeto	Texto	Deterioro
Daguerrotipo	Imagen de cámara	Metal	Reflexión	Positivo	Monocromático	1839 - 1860			Neutro	Muy brillante		Positivo-negativo		Delaminación de plata, corrosión de cobre.
Ambrotipo	Imagen de cámara	Vidrio	Reflexión	Positivo	Monocromático	1851 - 1865			Café	Brillante		Luces lechosas		
Ferrotipo	Imagen de cámara	Metal	Reflexión	Positivo	Monocromático	1855 - 1890			Café	Semi mate		Magnético		Corrosión, faltantes, craqueladuras
Calotipo	Impresión	Papel	Reflexión	Positivo	Monocromático	1840 - 1920	Una	Fibras visibles	Cian	Mate				
Albúmina	Impresión	Papel	Reflexión	Positivo	Monocromático	1851 - 1890	Dos	Fibras visibles	Amarillo, Café rojizo	Semi mate	Textura del papel	Soporte primario delgado, Soporte secundario grueso		Craqueladuras amarillamiento perdida de densidad en las luces y sombras
Carbón	Impresión	Papel	Reflexión	Positivo	Monocromático	1860 - 1940	Dos		Otro		Relieve en sombras			
Colodión de impresión directa	Impresión	Papel	Reflexión	Positivo	Monocromático	1885 - 1910	Tres	No se ven fibras	Purpura_Rojo	Brillante	Textura lisa	Soporte primario grueso, Soporte secundario grueso, Iridiscencia		Abrasiones, perdida de densidad en las luces
Plata gelatina de impresión directa	Impresión	Papel	Reflexión	Positivo	Monocromático	1885 - 1910	Tres	No se ven fibras	Amarillo	Brillante	Textura lisa	Soporte primario grueso, Soporte secundario grueso, Iridiscencia		perdida de densidad en las luces
Goma bicromatada	Impresión	Papel	Reflexión	Positivo	Monocromático	1890 - 1930	Dos		Otro					Ghosting
Platinotipo	Impresión	Papel	Reflexión	Positivo	Monocromático	1890 - 1930	Una	Fibras visibles	Neutro					
Plata gelatina	Impresión	Papel	Reflexión	Positivo	Monocromático	1890 - 2018	Tres	No se ven fibras	Neutro					
Colodión mate de impresión directa	Impresión	Papel	Reflexión	Positivo	Monocromático	1895 - 1910	Tres	No se ven fibras	Neutro, Café (platino), purpura (oro)	Semi mate	Texturizado	Soporte primario grueso, Soporte secundario grueso y de color,		Ghosting, No hay perdida de densidad
Difusión de plata	Impresión	Papel	Reflexión	Positivo	Monocromático	1942 - 2018	Tres		Neutro		Lisa	Pestañas, borde irregular o perforado, superficie con restos de adhesivo,	Polaroid	Revelado irregular, amarillamiento y desvanecimiento por recubrimiento irregular. Craqueladuras
Cromógeno	Impresión	Papel	Reflexión	Positivo	Policromático	1940 - 2018	Tres			Brillante	Lisa o texturizada	Papel de fibra, RC; acetato pigmentado		<1980 amarillamiento, pérdida de fijación de color y desvanecimiento de colorantes
Difusión de colorantes por transferencia	Impresión	Papel	Reflexión	Positivo	Policromático	1963 - 2018	Tres			Brillante	Lisa	Pestañas, borde irregular o perforado, superficie con restos de adhesivo,	Polaroid	Revelado irregular,
Blanqueo de colorantes	Impresión	Papel	Reflexión	Positivo	Policromático	1963 - 2018	Tres							
Difusión de colorantes por transferencia interna	Impresión	Papel	Reflexión	Positivo	Policromático	1972 - 2018	Tres			Brillante	Lisa,	Marco blanco con contenedor de químicos de procesado	Polaroid	Revelado irregular, craqueladuras, migración de colorantes en areas blancas
Colodión húmedo	Negativo	Vidrio	Transmisión	Negativo	Monocromático	1851 - 1885			Café		Aglutinante irregular, Barniz	Vidrio grueso e irregular		Abrasión
Gelatina seca	Negativo	Vidrio	Transmisión	Negativo	Monocromático	1880 - 1925			Neutro		Aglutinante irregular	Vidrio delgado y liso		Espejo de plata
Plata gelatina Nitrato de celulosa	Negativo	Plástico	Transmisión	Negativo	Monocromático	1889 - 1950						Muescas "V"	Nitrate	Amarillamiento del soporte, deformación,
Plata gelatina Acetato de celulosa	Negativo	Plástico	Transmisión	Negativo	Monocromático	1935 - 2018						Muescas "U"	Safety	Canales, burbujas, deformación, olor a vinagre
Plata gelatina Poliester	Negativo	Plástico	Transmisión	Negativo	Monocromático	1955 - 2018						Birrefringencia	Safety	
Cromógeno Acetato de celulosa	Negativo	Plástico	Transmisión	Negativo	Policromático	1942 - 2018							Safety	
Cromógeno Poliester	Negativo	Plástico	Transmisión	Negativo	Policromático	1955 - 2018						Birrefringencia	Safety	
Plata gelatina Vidrio	Transparencia	Vidrio	Transmisión	Positivo	Monocromático	1880 - 1940								
Plata gelatina Acetato de celulosa	Transparencia	Plástico	Transmisión	Positivo	Monocromático	1935 - 2018							Safety	
Plata gelatina poliester	Transparencia	Plástico	Transmisión	Positivo	Monocromático	1955 - 2018						Birrefringencia	Safety	
Procesos aditivos	Transparencia	Vidrio	Transmisión	Positivo	Policromático	1907 - 1935		Retícula			Líneas rectas paralelas	Bajo brillo y saturación		Delaminación, puntos verdes
Cromógeno Acetato de celulosa	Transparencia	Plástico	Transmisión	Positivo	Policromático	1935 - 2018							Safety	
Cromógeno Poliester	Transparencia	Plástico	Transmisión	Positivo	Policromático	1955 - 2018						Birrefringencia	Safety	

Figura 6. Mismo conjunto de datos en diferentes formatos, de izquierda a derecha se muestra una hoja de cálculo, un archivo CSV y un archivo ARFF.



Preparación de los datos

De las 27 columnas originales de la hoja de cálculo, se mantienen 15 columnas o atributos que representan conjuntos de información lo más acotados posibles en un rango de valores bien identificado. Las razones principales para descartar algunas de las columnas son la ausencia de información en campos opcionales (por ejemplo, subtipo de soporte), y la dispersión de datos en el caso de las columnas donde se escriben comentarios o anotaciones que suelen ser únicos (por ejemplo, particularidades del objeto fotográfico).

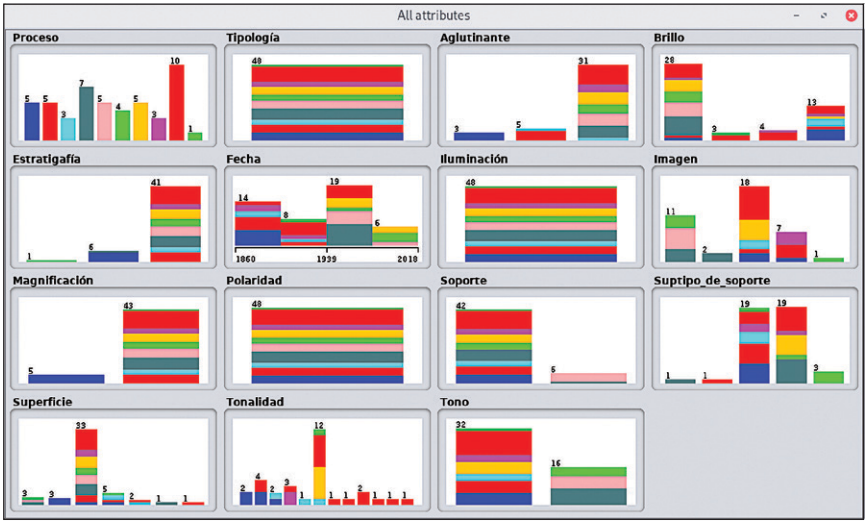
El método de almacenamiento nativo de WEKA es a través de archivos en formato ARFF. Este formato es muy similar a un archivo con formato de valores separado por coma (CSV), con la diferencia de que se le agregan encabezados que dan cuenta de las columnas o atributos, así como de los valores que pueden tener.

Este proceso de conversión de formato es simple de llevar a cabo gracias a que diversos programas permiten exportar bases de datos u hojas de cálculo en el formato CSV. Sin embargo, esta tarea se dificulta si los datos no están estandarizados, contienen errores ortográficos u otras inconsistencias. Por tal razón, la importancia de definir un vocabulario común es vital para un correcto funcionamiento del algoritmo y a su vez, para mantener consistencia e integridad de los datos.

Ejecución del algoritmo

Para utilizar el algoritmo C4.5 en WEKA, se hace uso de la versión de código abierto llamada algoritmo J4.8 (versión ligeramente mejorada del algoritmo C4.5 revisión 8 que está implementando en el lenguaje de programación Java). Con esto en consideración, el procedimiento es bastante directo, el primer paso consiste en seleccionar el archivo ARFF desde el explorador de WEKA para apreciar el conjunto de datos en su totalidad.

Figura 7. Gráficas de todas las propiedades a utilizar. Cada color representa un proceso fotográfico diferente.



En seguida se selecciona la pestaña de clasificación y se selecciona de la lista de opciones el algoritmo J48, se asignan los parámetros correspondientes, se selecciona el atributo o columna que representa las clases (en este caso el proceso fotográfico), se ejecuta el algoritmo y se observan los resultados que se obtienen.

Es crucial determinar los parámetros adecuados para obtener el mejor árbol de decisión. Si bien la cantidad de parámetros es acotada, son suficientes para generar diversidad de resultados que solamente pueden ser considerados mejores o peores, en comparación con el porcentaje de registros clasificados correctamente, según un conjunto de prueba (el cual puede ser el mismo conjunto de datos).

Los parámetros más significativos empleados para lograr un árbol de decisión suficientemente compacto, pero que logre clasificar correctamente la mayor cantidad de elementos no siempre es evidente y requiere de experimentación y múltiples ejecuciones del algoritmo.

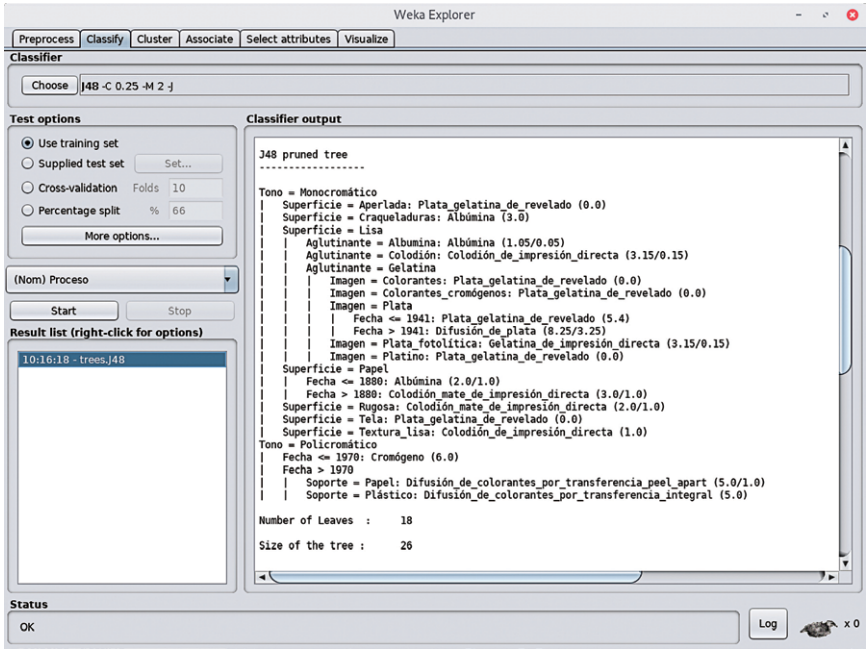
Tabla 4. Descripción de los parámetros relevantes en el algoritmo J4.8

Parámetro	Descripción	Valor usado
Unpruned tree (unpruned)	Usar árbol de decisión sin poda o recortes (genera árboles más grandes).	False
Reduced-error prunning (reducedErrorPrunning)	Determina si se utiliza un método alternativo de poda (no necesariamente con mejores resultados).	False
Pruning confidence (confidenceFactor)	Factor que se utiliza para podar o recortar el árbol; a menor valor mayor poda.	0.25
Minimum number of instances (minNumObj)	Establece el mínimo número de instancias por hoja o el número mínimo de ramificaciones.	2
MDL-correction (useMDLcorrection)	Ajuste basado en MDL (mínima longitud de descripción) para calcular ganancia de información.	False

Resultados

El resultado es un árbol de 26 nodos, 25 ramas y 18 son nodos terminales u hojas, que permiten clasificar fotografías según los 15 atributos que se definieron en la base de datos. Se tiene un porcentaje de instancias clasificadas correctamente de 85.41%, que es el porcentaje más alto que se alcanzó con los parámetros que ya se mencionaron. Esto no representa un árbol ideal que clasifique correctamente el 100% de las muestras pero da cuenta de una estructura diferente a la concebida de manera no-automatizada y hace evidente, al recorrer el árbol desde la raíz hasta una hoja, que hay características de la fotografía que tienen mayor peso para lograr una rápida clasificación de la misma.

Figura 8. Resultados mostrados en WEKA. Se aprecia una versión textual del árbol de decisión generado.

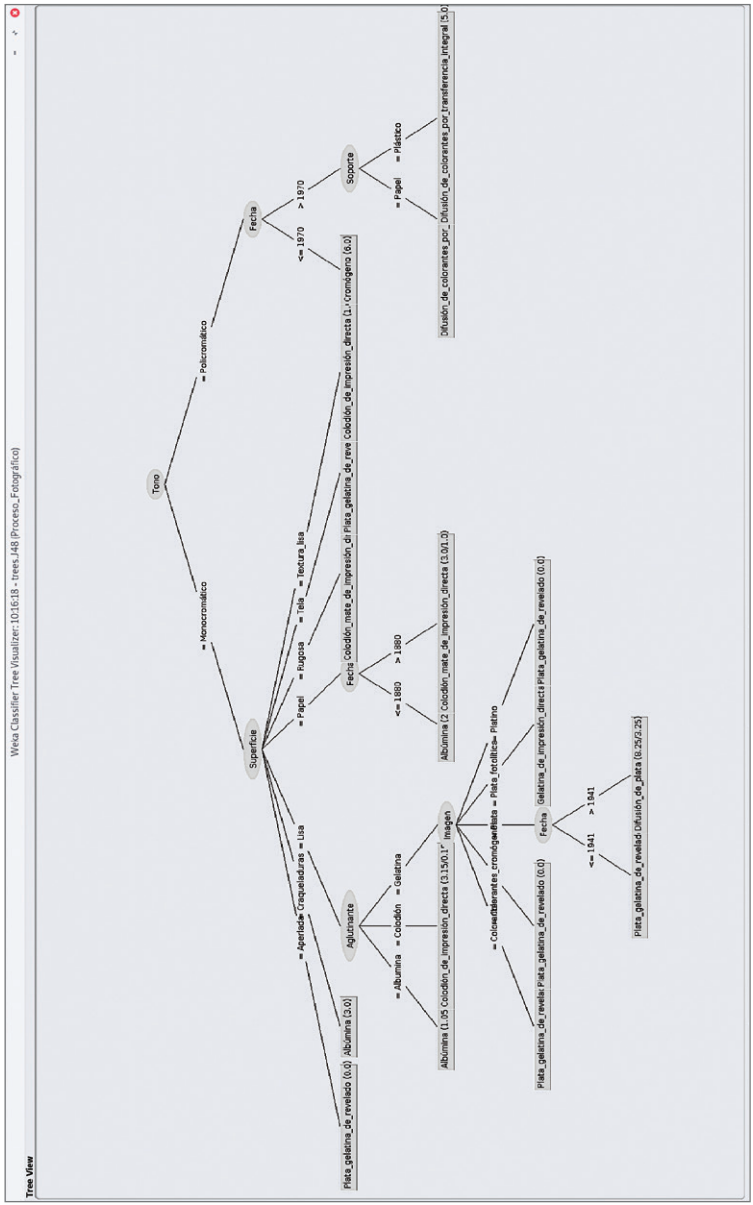


Conclusiones

Los conocimientos, experiencia y habilidades de especialistas, combinados con técnicas computacionales, nos permitieron obtener una metodología más robusta que cualquiera de estas dos opciones por separado. Se evidenciaron las diferencias entre las posibles maneras de construir un árbol de decisión y se revaloró la importancia de algunas características físicas de la fotografía. Una de las finalidades, buscar un balance entre un árbol compacto autogenerado y uno creado, a partir de la experiencia profesional, a fin de enriquecer también la práctica docente y de investigación.

A mayor cantidad de datos disponibles, el árbol de decisiones producido de manera automática, genera un modelo más refinado para contemplar diversos casos y mejorar la clasificación. Pero se debe

Figura 9. Representación gráfica del árbol de decisión final generado en WEKA.



tener cuidado con el problema del sobreajuste de datos, para lo cual se recomienda que los datos sean suficientemente representativos de las características físicas en las fotografías. Por lo tanto, no se debe descartar la opción de eliminar o agregar atributos y columnas en futuros procedimientos, ya que la experimentación y la comparación de diferentes versiones de árboles de decisión es de suma importancia en la labor comparativa.

Futuros trabajos encaminados en la misma temática pueden llegar a considerar una interfaz web a la que se le suministre el resultado de la generación automática de árboles de decisión, para mejorar la experiencia de usuario, e identificar de manera más rápida el proceso fotográfico, sin dejar de lado la parte didáctica con las explicaciones que ya se tienen. Así también, incluir otros tipos de análisis computacionales, como el uso de otras técnicas de Machine Learning o de análisis digital de la imagen asistida con herramientas de inteligencia artificial.

Bibliografía

Aasbø, Kristin, y Edwin Klijn. (2003). Sepiades: recommendations for cataloguing photographic collections : advisory report by the SEPIA Working Group on Descriptive Models for Photographic Collections. Amsterdam: European Commission on Preservation and Access.

Barra Moulain, Paula Alicia e Ignacio Gutiérrez Rubalcava. (2005). Normas catalográficas del Sistema Nacional de Fototecas del INAH. México: Instituto Nacional de Antropología e Historia.

Bell, Jason. (2015). *Machine Learning: Hands-On for Developers and Technical Professionals.* [S.l.]: John Wiley.

Boadas i Raset, Joan, Lluís-Esteve Casellas y M. Àngels Suquet i Fontana. (2001). Manual para la gestión de fondos y colecciones fotográficas. Girona: CCG Ediciones.

Comité Técnico de Normalización Nacional de Documentación. (2016). Norma Mexicana NMX-R-069-SCFI-2016. Documentos fotográficos. Lineamientos para su Catalogación. México: Secretaría de Economía.

Witten, Ian H., y Eibe Frank. (2016). *Data Mining: Practical Machine Learning Tools and Techniques, Fourth Edition.* San Diego: Elsevier Science & Technology Books.

Géron, Aurélien. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow. Concepts, Tools, and Techniques to Build Intelligent Systems.* 1st ed. Sebastopol, CA: O'Reilly.

_____. (2019). *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems.*

Heredia Herrera, Antonia. (1995). *Archivística general: teoría y práctica.* Sevilla: Diputación Provincial.

Image Permanence Institute. Graphics Atlas: Search Process. http://www.graphicsatlas.org/identification/.

Quinlan, Ross. (1986). *Induction of Decision Trees.* Machine Learning 1. 81-106. doi: https://doi.org/ 10.1007 /BF00116251.

Reilly, James. (1986). *Care and identification of 19th-century photographic prints.* Rochester: Eastman Kodak Company.

Inteligencia artificial y datos masivos en archivos digitales sonoros y audiovisuales. Instituto de Investigaciones Bibliotecológicas y de la Información/UNAM. La edición consta de 100 ejemplares. Coordinación editorial, Israel Chávez Reséndiz; revisión e specializada, Angélica Valenzuela y Valeria Guzmán González; revisión de pruebas, Valeria Guzmán González; formación editorial, Oscar Fernando Arcos Casañas. Fue impreso en papel cultural de 90 gr. en los talleres de Grupo Fogra. Año de Juárez 223. Col. Granjas San Antonio. Alcaldía Iztapalapa. Ciudad de México. Se terminó de imprimir en 2020.