VOLUMEN 21
XXVIII Verano De la Ciencia
ISSN 2395-9797
www.jóvenesenlaciencia.ugto.mx

# Clasificación temática y automática de imágenes de la red social Pinterest con deep learning

Aldo Isaac Hernández Antonio [1], Ana María Lizanette Becerra Cortés[1], Braulio José Baca Barbosa[1], Diana Martínez Frías [1], Diego Armando Gutiérrez Ayala[1], Mariana Esmeralda Centeno Reyes [1], Juan Carlos Gómez Carranza [1]*
[1]Departamento de Ingeniería Electrónica, División de Ingenierías Campus Irapuato-Salamanca, Universidad de Guanajuato.
{ai.hernandezantonio, aml.becerracortes, bj.bacabarbosa, d.martinezfrias, da.gutierrez.ayala, me.centenoreyes, jc.gomez}@ugto.mx

*Autor de correspondencia

## Resumen

La clasificación temática de imágenes consiste en asignar de manera automática una o más categorías predefinidas a una imagen dada. La clasificación de imágenes permite ajustar el contenido proporcionado a los usuarios con fines de marketing, información, entretenimiento, y más. En este artículo se presenta un proyecto de clasificación temática y automática de imágenes de la red social Pinterest. El objetivo del proyecto es asignar la categoría más adecuada a cada imagen de un conjunto de prueba, considerando 32 categorías predefinidas por el sitio. El proceso se divide en tres fases: la recolección de datos del sitio web de Pinterest, la extracción de características de las imágenes mediante técnicas de aprendizaje profundo (deep learning), y la clasificación de las imágenes utilizando modelos de aprendizaje de máquina (machine learning). Para extraer las características de las imágenes se utilizaron los modelos EfficientNetV2 y ConvNeXt con diferentes configuraciones. Para clasificar las imágenes en las categorías predefinidas, se aplicaron los modelos de k-vecinos más cercanos, máquinas de vectores de soporte, y una red neuronal de 3 capas. Los modelos de clasificación utilizaron las características extraídas en la segunda fase y fueron entrenados para predecir la categoría más adecuada para cada imagen. Para experimentar con los modelos mencionados, se utilizó un conjunto de datos compuesto por 32,000 imágenes obtenidas directamente del sitio web de Pinterest, las cuales corresponden a 670 usuarios y abarcan las 32 categorías establecidas. La evaluación de los modelos de clasificación se realizó utilizando la métrica macro F1.

**Palabras clave**: Clasificación de imágenes, aprendizaje de máquina, deep learning, redes sociales.

## 1. Introducción

La clasificación temática de imágenes es un proceso fundamental en el análisis y organización del contenido visual generado por los usuarios en plataformas en línea y redes sociales. La clasificación temática de imágenes en redes sociales tiene múltiples aplicaciones, ya que permite segmentar el contenido visual en grupos específicos según sus características. Esto proporciona a los usuarios una experiencia más personalizada, al tiempo que resulta invaluable para empresas y organizaciones en áreas como el marketing, la información, la educación y el entretenimiento, ya que pueden adaptar el contenido que ofrecen a sus audiencias de manera más efectiva. Por ejemplo, en el ámbito del marketing, la clasificación temática puede respaldar la creación de campañas enfocadas en productos o servicios dirigidos a audiencias específicas. Además, en términos de seguridad, la clasificación temática puede utilizarse para identificar contenido anómalo en las redes sociales, brindando una capa adicional de protección a los usuarios.

En el contexto de este proyecto, la clasificación temática se refiere a la tarea de determinar automáticamente la categoría más adecuada para una imagen que un usuario suba al sitio de la red social Pinterest; considerando una lista predefinida de 32 categorías propuestas por el mismo sitio. Estas categorías están definidas de acuerdo con las temáticas más relevantes en la plataforma. Para determinar la categoría más adecuada, se entrenan modelos de clasificación basados en aprendizaje automático. Estos modelos a su vez utilizan diferentes características visuales, o representaciones, de las imágenes obtenidas a través del uso de modelos de aprendizaje profundos.

El proyecto se dividió entonces en tres fases:

En la primera fase, se recopiló un conjunto de 1,069,477 imágenes del sitio web de Pinterest correspondientes a 670 usuarios, y organizadas en diferentes tableros [10]. Las imágenes estaban organizadas en 34 categorías definidas por Pinterest. Este conjunto de imágenes fue limpiado para quitar imágenes no legibles y aquellas pertenecientes a las categorías *none* y *other* (consideradas como no válidas), quedando 32 categorías con contenido. Finalmente, se seleccionaron al azar 1000 imágenes de cada categoría para realizar los experimentos, quedando un total de 32,000 imágenes.

En la segunda fase, se procesaron las imágenes obtenidas utilizando modelos de aprendizaje profundo (deep learning) para transformar cada imagen y representarla como un vector numérico de características visuales. Se utilizaron los modelos EfficientNetV2 y ConvNeXt con tres diferentes configuraciones cada uno, para un total de seis modelos de representación.

En la tercera fase del proyecto, se utilizó un enfoque de aprendizaje de máquina (machine learning) para construir modelos de clasificación utilizando las imágenes de un conjunto de entrenamiento, y posteriormente probar esos modelos para identificar la categoría más adecuada para cada imagen de un conjunto de prueba. En esta fase se implementaron los modelos de k-vecinos más cercanos (KNN), máquinas de vectores de soporte lineales (LSVM), y una red neuronal de 3 capas (NN). Es de esperar que los modelos aprendan a asociar las características de las imágenes a las categorías correspondientes, de manera que puedan realizar predicciones adecuadas sobre la categoría de una imagen desconocida. El desempeño de los modelos de clasificación se evaluó utilizando la métrica macro F1, que mide el desempeño por categoría.

La contribución de este trabajo radica en el análisis del desempeño de modelos de aprendizaje profundo y aprendizaje de máquina en conjunto para realizar la tarea de clasificación temática y automática de imágenes de Pinterest, de acuerdo con las categorías establecidas por el sitio. Se busca responder a las siguientes preguntas de investigación: 1) ¿Existe una arquitectura de aprendizaje profundo con un mejor rendimiento? 2) ¿Existe un modelo de aprendizaje de máquina con un mejor rendimiento? 3) ¿Existe una combinación de (modelo de aprendizaje profundo, modelo de aprendizaje de máquina) con un mejor rendimiento?

El resto del artículo está organizado de la siguiente manera. En la Sección 2 se presentará una breve descripción de trabajos relacionados con la clasificación temática de imágenes en redes sociales. En la Sección 3 se describirá detalladamente la metodología utilizada para abordar la tarea. La Sección 4 presentará los resultados obtenidos a partir de la experimentación con los modelos de aprendizaje profundo y de aprendizaje de máquina. Finalmente, en la Sección 5 se presentarán las conclusiones del trabajo y se plantearán posibles direcciones para futuras investigaciones.

## 2. Trabajos Relacionados

A lo largo de los años, la tarea de asignar automáticamente categorías a imágenes ha sido abordada mediante diversos enfoques dependiendo de las características de estas. Cuando las imágenes son acompañadas con un texto descriptivo o comentario sobre estas, este texto se puede analizar para obtener información semántica relacionada con el contenido visual y determinar la(s) categoría(s) más adecuada(s) para la(s) imágene(s). Sin embargo, lo más común es hacer un análisis directo de las imágenes para encontrar relaciones entre la información visual y las categorías que se quieren asignar.

En el trabajo realizado en [1], los autores exploraron el uso de los histogramas de color en combinación con características adaptativas de textura para representar las imágenes. A partir de esta combinación de características, los autores emplearon un clasificador de máquinas de vectores de soporte (SVM) para asociar las características de color y textura de las imágenes con las palabras clave asignadas al conjunto de imágenes de prueba.

Por otro lado, en [2] se utilizaron las características SIFT (Scale-Invariant Feature Transform), luminosidad, color y forma para representar las imágenes. Con esta combinación de características, se utilizó un modelo de clasificación basado en KNN para asignar las categorías a las imágenes de prueba.

En [3], los autores utilizan un modelo llamado User Image Latent Space Model para describir las imágenes, el cual agrupa los pixeles de la imagen de una manera jerárquica en palabras visuales, regiones semánticas y temas. Este modelo es utilizado en combinación con la distancia euclidiana para determinar la categoría de un conjunto de imágenes de prueba de la red social Flickr, utilizando categorías predefinidas por los autores.

En [4], los autores utilizaron una red neuronal convolucional (CNN) preentrenada para transformar las imágenes en vectores numéricos de características profundas. Posteriormente, los autores emplearon esas características para entrenar un modelo de propagación de etiquetas a nivel de imagen y de grupo, el cual fue utilizado para predecir los intereses (expresados como categorías) de un grupo de usuarios de la red social Pinterest.

En ese sentido, el uso de los modelos de redes neuronales profundas, tales como los CNN, ha cobrado relevancia en los últimos años cuando se trabaja en problemas de análisis de imágenes, incluyendo la transformación de las imágenes y su clasificación [5, 6]. Este tipo de redes neuronales han demostrado excelentes rendimientos en importantes tareas de clasificación de imágenes y detección de objetos [7, 8, 9].

Se puede decir que los avances del etiquetado automático de imágenes han contribuido significativamente a mejorar la capacidad de comprensión y clasificación de contenido visual, teniendo un impacto positivo en una amplia gama de aplicaciones y campos de estudio.

# 3. Metodología

La metodología se divide en tres fases, tal como se muestra en la Figura 3.1. Estas fases se describen en las siguientes subsecciones.

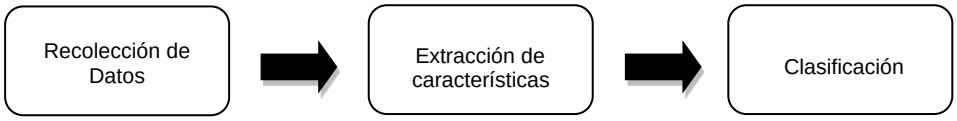



*Figura 3.1. Fases que conforman la metodología*

## 3.1 Recolección de datos

Para este artículo se utilizó un conjunto de imágenes el cual fue recopilado previamente del sitio web de Pinterest [10]. Siguiendo la estructura de Pinterest, que organiza las imágenes por usuario, para cada usuario hay un grupo de tableros y en cada tablero un grupo de imágenes, los datos se encuentran organizados en directorios, extrayendo de cada usuario todos sus tableros y los pines correspondientes a cada tablero. En la Figura 3.2 hay 5 paneles, que ilustran un ejemplo de la organización de los datos. El primer panel corresponde al nombre de los usuarios, en el segundo se encuentra la información general sobre los tableros de un usuario en particular (7crazycats), en el tercer panel están los tableros de ese usuario, en el cuarto están las carpetas que representan cada uno de los pines en un tablero (april-fool-s) y en el último se puede ver la información de un solo pin (225813368786172656).

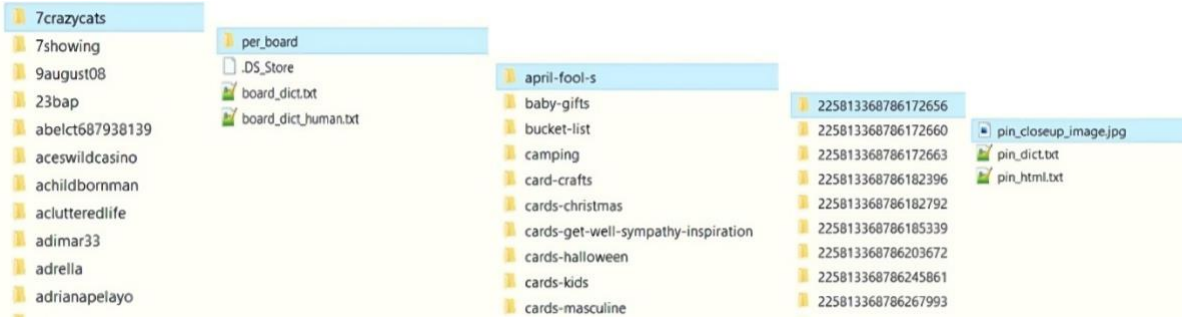



*Figura 3.2. Ejemplo del directorio de un usuario*

### 3.1.1. Descripción del conjunto de datos

El conjunto de imágenes recopilado contiene información de 670 usuarios, cada usuario tiene un número diferente de tableros y, cada uno de estos tableros tiene un número variable de imágenes, resultando un conjunto total de 1,069,477 imágenes.

Pinterest proporciona 34 categorías predefinidas para clasificar las imágenes que se suben a su sitio. Estas categorías se muestran en la Tabla 3.1. Cuando un usuario crea un tablero, Pinterest le permite seleccionar una de las categorías para describirlo de manera general. La categoría asignada al tablero define la categoría para todas las imágenes dentro de él.

| | | | |
|---|---|---|---|
| animals | architecture | art | cars motorcycles |
| celebrities | design | diy crafts | education |
| film music books | food drink | gardening | geek |
| hair beauty | health fitness | history | holidays events |
| home decor | humor | illustrations posters | kids |
| mens fashion | outdoors | photography | products |
| quotes | science nature | sports | tattoos |
| technology | travel | weddings | womens fashion |
| *none* | *other* | | |

***Tabla 3.1.*** *Categorías predefinidas por Pinterest*

Del conjunto de imágenes utilizados en este trabajo, se excluyeron aquellas imágenes que no fueran legibles por algún problema con el archivo, y aquellas asociadas con las categorías *other* o *none*. Se eliminaron estas categorías porque no tienen una clasificación temática precisa, es decir, cualquier imagen puede relacionarse con alguna de esas dos categorías, lo que dificulta que un modelo de clasificación pueda aprender a discriminar con respecto a esas dos categorías. Después del proceso de limpieza, se obtuvieron un total de 598,279 imágenes, las cuales quedaron distribuidas por categoría como se muestra en la Figura 3.3. En esta figura se observa que la categoría *womens_fashion* es la más popular, mientras que la categoría *geek* es la menos popular. Finalmente, del total de imágenes, se seleccionaron de forma aleatoria 1,000 imágenes para cada categoría, quedando un conjunto de 32,000 imágenes con el que se realizaron los experimentos.

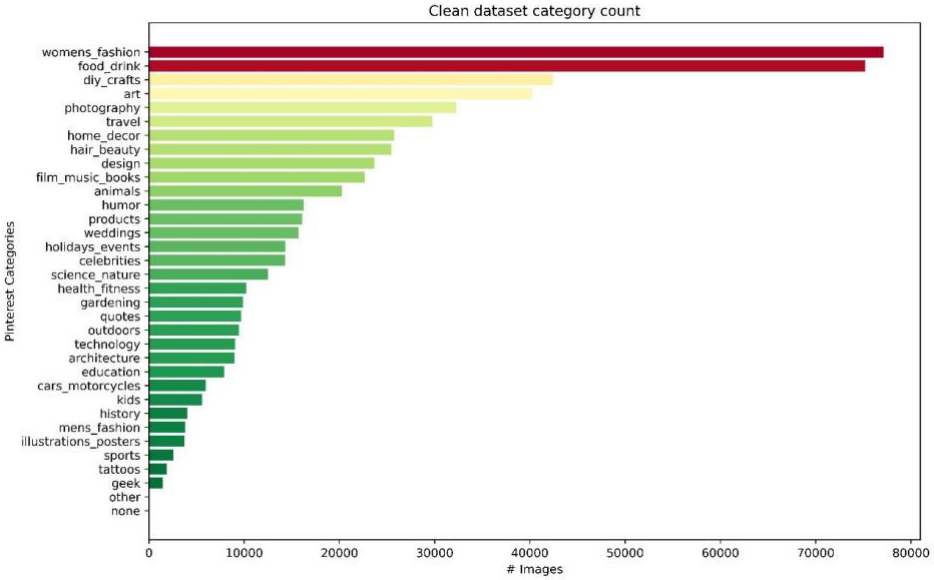



*Figura 3.3. Distribución de imágenes por categoría*

## 3.2. Extracción de características

Las imágenes seleccionadas se transformaron utilizando modelos de aprendizaje profundo basados en redes neuronales convolucionales. En particular se utilizaron los modelos EfficientNetV2 [11], en sus configuraciones B0, B1 y B3; y ConvNext [12], en sus configuraciones Tiny, S y B. Cada modelo de red produce como salida para cada imagen un vector numérico que representa una serie de características visuales que describen a la imagen de acuerdo con la configuración de la red. Cada configuración de un modelo de red cambia el número de neuronas en las diferentes capas de la red, lo que puede producir diferentes tamaños en los vectores de salida. Particularmente, en este proyecto se utilizaron versiones de los modelos que fueron preentrenados con el conjunto de datos de Imagenet [13]. En la Figura 3.4 se ilustra la arquitectura del modelo EfficientNetV2, de esta figura, la capa llamada *Fully Connected* es la que representa el vector de características que se extrae para cada imagen. En la Tabla 3.2 se muestra el tamaño de los vectores de características que produce cada uno de los modelos de red utilizados en este trabajo.

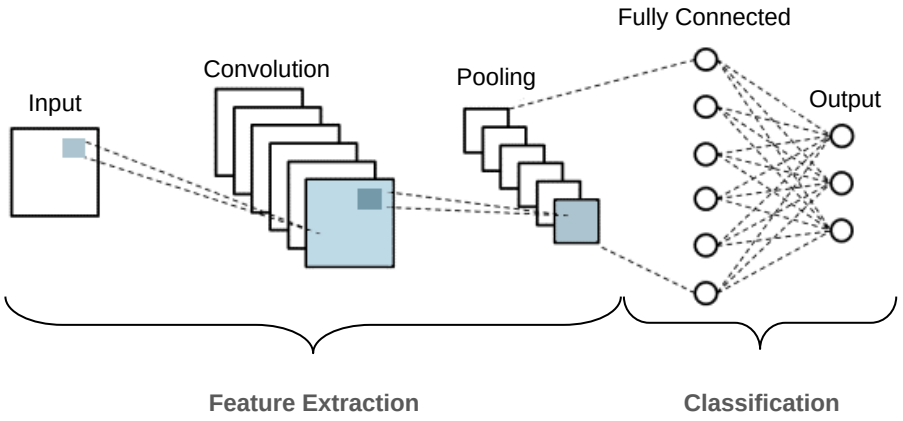



*Figura 3.4. Representación del modelo EfficientNetV2*

| EfficientNetV2 | | | ConvNext | | |
|---|---|---|---|---|---|
| **B0** | **B1** | **B3** | **Tiny** | **B** | **S** |
| 1280 | 1280 | 1536 | 768 | 1024 | 768 |

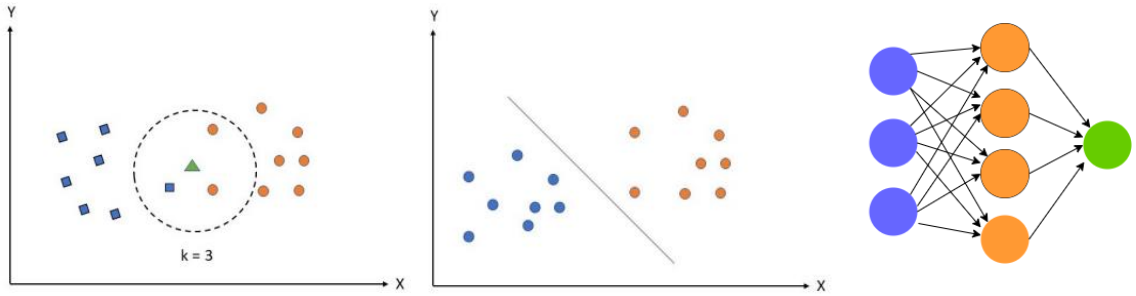
***Tabla 3.2.*** *Tamaño de los vectores de cada modelo*

El resultado del proceso de transformar las imágenes a vectores numéricos se almacenó en archivos para tener un procesamiento más eficiente en las fases posteriores.

## 3.3. Clasificación

Una vez procesadas las imágenes para representarlas como vectores numéricos, se procedió a la creación y prueba de modelos de clasificación. En esta fase se siguió un enfoque de aprendizaje de máquina, en donde las imágenes fueron separadas en dos: 80% (25,600 imágenes) se usaron para formar el conjunto de entrenamiento y 20% (6,400 imágenes) se usaron para formar el conjunto de prueba. La separación se hizo de forma estratificada, de tal forma que cada categoría en ambos conjuntos tenía el mismo número de imágenes. Usando el conjunto de entrenamiento se construyeron modelos de aprendizaje. Estos modelos pueden posteriormente clasificar de manera automática imágenes no vistas en las categorías definidas.

Para este trabajo se utilizaron los modelos de k-vecinos más cercanos (KNN), máquinas de vectores de soporte lineales (LSVM), y una red neuronal de 3 capas (NN). KNN es un modelo basado en instancias que clasificará a una imagen de acuerdo con la similitud de K otras imágenes en el conjunto de entrenamiento. LSVM es un modelo discriminativo, que forma un hiperplano para separar las categorías. En el caso de que haya varias categorías, se forman n hiperplanos (n = número de categorías) en un proceso *one-vs-all* (una categoría contra todas las demás). NN es un modelo similar a LSVM, pero el hiperplano no es lineal sino una curva que permite la separación de todas las clases al mismo tiempo. La Figura 3.4 muestra una representación visual de cada modelo de clasificación.




a) b) c)

*Figura 3.4. Representación de los modelos a) KNN, b) LSVM, c) Red Neuronal*

Durante el proceso de entrenamiento, los modelos de clasificación aprenden a mapear las características visuales encontradas en los vectores numéricos que representan las imágenes a las categorías correspondientes. De esta forma, pueden posteriormente realizar predicciones sobre la categoría de una imagen desconocida. Para esto, una vez construidos los modelos, se utilizó el conjunto de prueba para probar qué también hacen la tarea de clasificación.

Considerando las imágenes del conjunto de prueba, para cada categoría cada imagen podría clasificarse con dos opciones: pertenece a la categoría (considerada como positiva) o no pertenece a la categoría (considerada como negativa). Esto da como resultado una matriz con cuatro celdas para cada categoría, también conocida como matriz de confusión. Esta matriz relaciona las categorías reales de las imágenes con las categorías predichas por un modelo. En la Figura 3.5 se observa la representación visual de la matriz de confusión, en donde, TP corresponde al número de verdadero positivos, TN al número de verdaderos negativos, FP al número de falsos positivos y FN al número de falsos negativos.

| | | Categoría predicha | |
|---|---|---|---|
| | | Positivo | Negativo |
| Categoría Real | Positivo | **TP** | **FN** |
| | Negativo | **FP** | **TN** |

*Figura 3.5. Representación de la matriz de confusión.*

Utilizando la matriz de confusión se pueden calcular varias métricas que indiquen el desempeño del modelo en la clasificación. En este trabajo se utilizó la métrica F1, la cual está basada en las métricas de *precision* y *recall*. La *precision* mide la proporción de imágenes clasificadas correctamente, es decir, se centra en lo que el modelo dice y luego lo compara con la realidad; mientras que *recall* mide cuántas imágenes positivas son correctamente clasificados. La métrica F1 toma valores entre 0 (clasificación totalmente errónea) y 1 (clasificación perfecta) y está definida por las siguientes ecuaciones:

$$Precision = \frac{TP}{TP + FP} \quad (3.1)$$

$$Recall = \frac{TP}{TP + FN} \quad (3.2)$$

$$F1 = 2\,\frac{Precision * Recall}{Precision + Recall} \quad (3.3)$$

En la versión macro de la métrica F1, ésta se calcula para cada categoría de manera independiente y al final se calcula el promedio de las métricas independientes. Esto ayuda a mitigar el efecto en el caso de que haya categorías que dominen por la cantidad de imágenes, al ser pesadas igual que una categoría con pocas imágenes.

Adicionalmente, cada modelo de clasificación tiene algunos hiperparámetros que influyen en su comportamiento al momento de realizar las predicciones, por lo que es recomendable tratar de encontrar los valores óptimos de esos hiperparámetros para el conjunto de datos con el que se trabaja. En este caso, se realizó una búsqueda en malla (grid search) utilizando una validación cruzada de cinco partes con el conjunto de entrenamiento. Básicamente, los datos de entrenamiento se dividieron en cinco partes, cada parte se utilizó una vez como conjunto de prueba y las cuatro partes restantes como conjunto de entrenamiento. El modelo es entrenado, probado y evaluado cinco veces, y al final se calcula el promedio de su desempeño. Este proceso se repite con distintos valores para los hiperparámetros. Aquel valor que dé mejores resultados se selecciona y se realiza el entrenamiento y prueba final, con los conjuntos de entrenamiento y prueba originales. Los hiperparámetros de cada modelo y sus valores utilizados en el proceso de optimización se muestran en la Tabla 3.3.

| Modelo | Parámetro | Descripción | Valores |
|---|---|---|---|
| **KNN** | k | Número de vecinos a considerar en la predicción | [1, 5, 10, 20, 50] |
| **LSVM** | C | Parámetro de regularización | [0.01, 0.1, 1, 10, 100] |
| **NN** | batch_size | Número de ejemplos que se propagan en la red | [32, 64, 128] |

***Tabla 3.3.*** *Hiperparámetros y sus valores utilizados en la optimización*

Todos los códigos para la transformación de las imágenes, entrenamiento y prueba de los modelos de clasificación se realizaron en Python utilizando las bibliotecas scikit-learn, pandas, numpy, Keras y TensorFlow.

# 4. Resultados

En la Tabla 6 se muestran los resultados para la métrica macro F1 de la clasificación del conjunto de imágenes de prueba con los diferentes modelos de clasificación, en combinación con las diferentes características para representar a las imágenes extraídas utilizando los modelos de aprendizaje profundo. En la última columna se muestra el promedio del desempeño de todos los modelos considerando una característica en particular. En el último renglón se muestra el promedio del desempeño de un modelo considerando todas las características. En negritas se muestran los valores más altos, tanto general como de los promedios.

| | Modelo de Clasificación | | | |
|---|---|---|---|---|
| **Características de aprendizaje profundo** | **KNN** | **LSVM** | **NN** | **Promedio** |
| **EfficientNetV2_B0** | 0.042 | 0.040 | 0.002 | 0.028 |
| **EfficientNetV2_B1** | 0.090 | 0.076 | 0.062 | 0.076 |
| **EfficientNetV2_B3** | 0.089 | 0.087 | 0.053 | 0.076 |
| **ConvNext_Tiny** | 0.077 | 0.083 | 0.067 | 0.076 |
| **ConvNext_S** | 0.077 | 0.086 | 0.075 | 0.079 |
| **ConvNext_B** | **0.115** | 0.110 | 0.088 | **0.104** |
| **Promedio** | **0.082** | 0.080 | 0.058 | |

***Tabla 4.1**. Resultados (macro F1) para los diferentes modelos de clasificación y de características de aprendizaje profundo*

En esta tabla es posible observar que los valores obtenidos son muy bajos, cercanos a 0. Esto es un indicativo de que la tarea de clasificación temática automática de imágenes en la red Pinterest es complicada. Los modelos no son capaces de identificar de manera efectiva una relación entre las características visuales de las imágenes y su correspondiente categoría. El valor más alto lo obtuvo el modelo KNN en combinación con las características ConvNext en su versión B, con un valor de 0.115. El valor más bajo fue para el modelo NN con las características EfficientNetV2 en su versión B0, con un valor de 0.002.

Revisando los promedios, es posible observar que el modelo de clasificación con el mejor desempeño considerando todas las características es KNN, mientras que el modelo con peor desempeño es NN. El modelo KNN funciona a través de medir la similitud entre el vector de una imagen de prueba y los vectores de las imágenes en el conjunto de entrenamiento. Este desempeño indica que tal enfoque es adecuado al momento de determinar la categoría de una imagen desconocida. Por otro lado, el modelo de NN utiliza un discriminador no lineal, el cual aparentemente no es capaz de separar de forma adecuada las imágenes entre sus diferentes categorías.

Por otro lado, las características con mejor desempeño considerando todos los modelos de clasificación son ConvNext en su versión B, mientras que las características con peor desempeño son EfficientNetV2 en su versión B0. Esto parece indicar que la arquitectura ConvNext obtiene características visuales que están más asociadas con las categorías de las imágenes.

# 5. Conclusiones

En este trabajo se presentó un estudio sobre la clasificación automática y temática de imágenes en la red social Pinterest. La tarea consistió en poder determinar la categoría más adecuada para una imagen, considerando un conjunto de 32 categorías posibles predefinidas por la propia red social. En el estudio se consideraron dos modelos de aprendizaje profundo, con tres variaciones en su configuración para cada uno, para un total de seis modelos, encargados de transformar las imágenes a vectores numéricos que representaban características visuales de las imágenes. Con estas características se entrenaron y probaron tres modelos de clasificación basados en aprendizaje de máquina para evaluar su desempeño en la tarea de la asignación de categorías. Para la experimentación se utilizó un conjunto de imágenes extraído directamente del sitito web de Pinterest compuesto por 32,000 imágenes.

De acuerdo con los resultados de los experimentos, es posible concluir lo siguiente:

1. El problema planteado en este trabajo es complicado de resolver. Los resultados obtenidos fueron muy bajos para los diferentes modelos de clasificación y características visuales extraídas, por lo

que se requiere seguir explorando más alternativas de solución.

2. Respondiendo a la primer pregunta planteada en la introducción, la arquitectura que produce un mejor desempeño en la clasificación es la ConvNext en su versión B, aunque el desempeño sigue siendo bajo.
3. Respondiendo a la segunda pregunta, el modelo de aprendizaje de máquina con un mejor desempeño es KNN, también con resultados aún bajos.
4. Respondiendo a la tercer pregunta, la combinación de modelo de aprendizaje profundo y de aprendizaje de máquina con un mejor desempeño fue KNN con ConvNext en su versión B.

Algunas ideas para trabajos futuros incluyen el uso de otras arquitecturas de redes neuronales de aprendizaje profundo para representar las imágenes, tales como ResNet, EfficientFormer, etc.; así como el uso de otros modelos de aprendizaje de máquina para la clasificación, como máquinas de vectores de soporte no lineales y ensambles de clasificadores. También es posible considerar el uso de una mayor cantidad de imágenes para construir los modelos de clasificación, lo cual se espera. Finalmente, la combinación de características o de la predicción de los modelos, es otra área para explorar.

## Bibliografía/Referencias

[1] Feng, H., Shi, R., & Chua, T. S. (2004, October). A bootstrapping framework for annotating and retrieving WWW images. In Proceedings of the 12th annual ACM international conference on Multimedia (pp. 960-967).

[2] Boiman, O., Shechtman, E., & Irani, M. (2008, June). In defense of nearest-neighbor based image classification. In 2008 IEEE conference on computer vision and pattern recognition (pp. 1-8). IEEE.

[3] Xie, P., Pei, Y., Xie, Y., & Xing, E. (2015, February). Mining user interests from personal photos. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 29, No. 1).

[4] You, Q., Bhatia, S., & Luo, J. (2016). A picture tells a thousand words—About you! User interest profiling from user generated visual content. Signal Processing, 124, 45-53.

[5] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition una ected by shift in position. Biological Cybernetics, 36:193-202, 1980.

[6] L. Bottou Y. Lecun, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, page 2278-2324, 1998.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. Curran Associates, Inc., page 1097-1105, 2012.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang,A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. International Journal of Computer Vision (IJCV), 2015.

[9] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. CoRR, vol. abs/1412.2306, 2014.

[10] Cinar, Y. G., Zoghbi, S., & Moens, M. F. (2015, November). Inferring user interests on social media from text and images. In 2015 IEEE international conference on data mining workshop (ICDMW) (pp. 1342-1347). IEEE.

[11] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.

[12] Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11976-11986).

[13] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.