Procesamiento de Imágenes y Extracción de Características Morfológicas para Clasificación de Galaxias

J.Z. Salinas, C.E. Schaerer, H. Legal
Facultad Politécnica, Universidad Nacional de Asunción
San Lorenzo, Paraguay, 2016

M. García-Torres
Universidad Pablo de Olavide, Escuela de Ingeniería
Sevilla, España

***Resumen*—En este TFG realizamos una comparación entre múltiples algoritmos de minería de datos al problema de clasificación morfológica de galaxias. Las imágenes son procesadas para estandarizarlas para eliminar los efectos de orientación y traslación. Extraemos de las imágenes estandarizadas, las características morfológicas, los componentes principales y los componentes independientes y comparamos diferentes algoritmos de aprendizaje automático supervisado con dichas características. Los experimentos muestran que los mejores resultados son alcanzados con el algoritmo *Naïve Bayes* y con los componentes principales como características.**

Palabras clave: Procesamiento de imágenes, minería de datos, aprendizaje automático, clasificación de galaxias.

I. Introducción

Las galaxias son un conjunto de estrellas, nubes de gas, planetas, polvo cósmico, materia oscura y energía que permanecen unidos mediante la acción de la fuerza de gravedad y aislados de sistemas similares por grandes regiones de espacio vacío. La apariencia visual de las galaxias proporciona a los astrónomos mucha información sobre la composición y evolución de las mismas. Esta apariencia está en función de la edad de las mismas, de los elementos que están compuestas y de su proceso de formación, por lo que las galaxias más jóvenes (más lejanas) todavía no tienen una estructura definida y son más pequeñas que las galaxias más viejas. Entre las galaxias más viejas, están las elípticas y las espirales, cuyo proceso de formación aún no se sabe con certeza, aunque existen teorías que explican dichas formas [1].

En este trabajo realizamos la clasificación de galaxias basándonos en la "Secuencia de *Hubble*", cuya clasificación se define a continuación:

La Secuencia de *Hubble*

Hubble ha basado su clasificación en fotografías de las galaxias tomadas con telescopios de la época (alrededor de 1936). Al principio creyó que las galaxias elípticas eran una forma inicial, y que posteriormente evolucionaban a espirales. Nuestro conocimiento actual sugiere que la situación podría ser opuesta, no obstante esta creencia quedó en la jerga de astrónomos que aún hablan de "tipo primitivo" o "tipo avanzado" cuando se refieren a galaxias elípticas y espirales, respectivamente. *Hubble* dividió los tipos de galaxias según la siguiente clasificación (**Figura 1**):

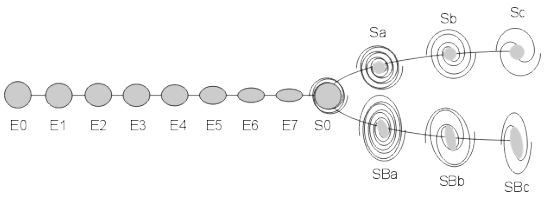

Figura 1: Secuencia de *Hubble*

- **Galaxias elípticas**: (E0-7) tienen forma elíptica, con una distribución bastante uniforme de las estrellas por todas partes. El número indica el grado de excentricidad: las galaxias E0 son casi redondas, mientras que las E7 son muy aplanadas. El número indica solo la apariencia de la galaxia en el cielo, no su geometría real.
- **Galaxias lenticulares**: (S0 y SB0) parecen tener una estructura de disco con una concentración de estrellas central proyectándose de él. No muestran ninguna estructura espiral.
- **Galaxias espirales**: (Sa-d) tienen una concentración de estrellas central y un disco aislado que presenta brazos espirales. Los brazos están centrados alrededor de la protuberancia, variando de los muy arremolinados y poco definidos (Sa) a los muy sueltos y definidos (Sc y Sd). Así, mientras que en las primeras, la concentración central es muy pronunciada, en las últimas lo es bastante menos, y salvo excepciones, la cantidad de estrellas jóvenes y la proporción de gas van aumentando a lo largo de la secuencia.
- **Galaxias espirales barradas**: (SB0/a-d) tienen una estructura en espiral, similar a las galaxias espirales pero los brazos se proyectan desde el final de una "barra" central en lugar de emerger de una concentración central, como cintas en los extremos de una vara. De nuevo, SBa a SBd indica como de arremolinados están estos brazos y el grado de desarrollo de la concentración central, y de nuevo, salvo excepciones, al ir progresando en la secuencia, la cantidad de gas y estrellas jóvenes va aumentando.
- **Galaxias espirales intermedias**: (SAB0/a-c) tienen una morfología intermedia entre las galaxias espirales y las

galaxias espirales barradas.
- **Galaxias irregulares**: (Irr) se dividen en Irr-I, que muestran estructura espiral deformada, e Irr-II para las galaxias que no encajan en ninguna otra categoría.

II. Definición del Problema

Antes de la era de la información, las observaciones se hacían solo mirando a través del ocular de los telescopios o analizando con una lupa las placas fotográficas, por lo que las clasificaciones y categorizaciones de objetos celestes se realizaban de forma manual y lineal. En otras palabras, la categorización de los objetos se realizaba uno a uno.

Con la llegada de las tecnologías de información, la astronomía tuvo una rápida evolución hacia la automatización. La captura de datos ya no era un trabajo de "uno a uno" sino que empezaron a llegar grandes cantidades de datos de cientos y miles de objetos simultáneamente, tanto de distintos observatorios como de distintas fuentes (luz visible, rayos X, ultravioleta, rayos gamma, radiofrecuencias, entre otros).

Con esto, empieza un nuevo problema, y es la capacidad para analizar toda esta información en un tiempo razonable. La clasificación manual ya no es una tarea viable por lo que se necesitan de alternativas para lidiar con esta situación.

III. Fases de un Sistema de Clasificación de Imágenes de Galaxias

La clasificación morfológica automática de galaxias ya lleva como dos décadas de estudio. Existen varias propuestas pero todas se basan en un mismo esquema de procesos como se muestra a continuación:

a - **Preprocesamiento de Imágenes**: incluye generalmente un proceso de filtrado para eliminación de ruido (variación aleatoria de brillo o color que no corresponde con la realidad) y técnicas para realzar los detalles importantes y estandarizar el formato de las imágenes.

b - **Extracción de Características**: se logra identificando características asociadas a la morfología del objeto. Entre las características podemos citar [2]: elongación, factor de forma, convexidad, factor de forma rectangular e índice de asimetría. También existen otros tipos de características (no morfológicas) como lo son las obtenidas con el PCA [2], [3].

c - **Proceso de Clasificación**: se logra utilizando técnicas de minería de datos junto con las características extraídas de las galaxias para identificar a qué clase o grupo pertenecen.

III-A. Preprocesamiento de Imágenes

Las imágenes de galaxias generalmente difieren en tamaño, colores, formato, orientación y en la mayoría de las veces, la galaxia contenida en la imagen no está centrada. Para evitar esta heterogeneidad, en esta fase creamos imágenes invariantes al color, orientación y posición. Primero, aplicamos un filtro RGB a escala de grises mediante la siguiente ecuación:

$$y = 0{,}587\, g + 0{,}299\, r + 0{,}114\, b \tag{1}$$

donde g, r y b representan los colores verde, rojo y azul respectivamente, y es la intensidad del pixel en escala de grises siguiendo la recomendación del estándar 601 de la Unión Internacional de Telecomunicaciones [4]. En la **Figura 2** podemos ver algunos ejemplos de imágenes de distintos tipos de galaxias en diferentes posiciones que son obtenidas en escala de grises mediante la ecuación (1) desde una base de datos de un observatorio virtual.

Figura 2: Imágenes en escala de grises

Segundo, un umbral determinado es aplicado para binarizar la imagen y remover ruido con (2)

$$B(i,j) = \begin{cases} 1 & \text{si} \quad I(i,j) > \tau \\ 0 & \text{otro caso} \end{cases} \tag{2}$$

donde I es la imagen original (**Figura 2**), B la imagen binarizada, τ es el umbral y los índices i y j representan las filas y columnas de los pixeles de la imagen. En este TFG asignamos a τ el valor 50. En la **Figura 3** podemos ver el resultado de aplicar binarización a la imagen original.

Figura 3: Imágenes binarizadas

Tercero, un filtro de apertura es aplicado a la imagen binarizada para remover el ruido restante con (3)

$$A = B \circ E = (B \ominus E) \oplus E \tag{3}$$

Figura 4: Imágenes filtradas

donde B es la imagen binarizada, E es el elemento estructurante y los operadores $\ominus$ y $\oplus$ representan erosión y dilatación respectivamente. En este TFG aplicamos el filtro de apertura utilizando una matriz de unos de tamaño 2 x 2 como elemento estructurante y realizando una sola iteración. En la **Figura 4** vemos las imágenes luego de realizar un filtro de apertura para eliminar la mayor cantidad de ruido posible.

Cuarto, los pixeles de la imagen original que corresponden con los pixeles de la imagen con mayor contorno obtenida con la ecuación (3) es extraída con (4)

$$Y = I \cap A \tag{4}$$

donde I es la imagen original, A el resultado de hacer apertura con (3) y el operador $\cap$ representa la intersección. En la **Figura 5** vemos el resultado de interceptar la imagen original con la imagen filtrada utilizada como máscara.

Figura 5: Región de interés extraída

Quinto, la galaxia es centrada al recuadro de la imagen con una transformación afín (5), con la matriz de transformación **M** como se muestra en (6)

$$z = \mathbf{M}x + w \tag{5}$$

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \end{bmatrix}, \quad \begin{cases} t_x = cg_x - ci_x, \\ t_y = cg_y - ci_y \end{cases} \tag{6}$$

donde z es la imagen transformada, w corresponde a la traslación, $\mathbf{M}$ corresponde a los cambios de escala, rotaciones y sesgos, cg y ci son el centro de la galaxia y el centro del recuadro de la imagen respectivamente. En la **Figura 6** vemos las galaxias trasladadas al centro del cuadro.

Figura 6: Imágenes trasladadas al centro

Finalmente, la galaxia es rotada para alinear el eje principal con el eje horizontal, el ángulo del eje principal es calculado con (7)

$$\alpha = \arctan(a_x/a_y) \tag{7}$$

donde a_x y a_y son los valores de fila y columna del eje principal. En la **Figura 7** vemos las galaxias rotadas con el eje mayor alineado con la horizontal.

Figura 7: Imágenes alineadas con la horizontal

III-B. Extracción de Características

En esta fase, varias características morfológicas y no morfológicas son extraídas de las imágenes procesadas.

III-B1. Características Morfológicas: Las MFs están basadas en la percepción visual de las galaxias. Los resultados experimentales obtenidos en [2] demuestran que las MFs son más efectivas que los PCs. Por esta razón, usaremos ICA como extractor de características para verificar si los ICs tienen un mejor desempeño. A continuación se definen las fórmulas utilizadas para la extracción de MFs.

La elongación (e) es definida como:

$$e = \frac{(a-b)}{(a+b)} \tag{8}$$

donde a y b son los ejes mayor y menor de la elipse que envuelve a la galaxia.

El factor de forma (F_f) es definido como:

$$F_f = \frac{A(I)}{P(I)} \tag{9}$$

donde $A(I)$ es el área de la imagen I y $P(I)$ es el perímetro de la imagen I.

La convexidad (C_v) está definida por:

$$C_v = \frac{P(I)}{P(B_r(I))} \tag{10}$$

donde $B_r(I)$ es el mínimo rectángulo que envuelve a la galaxia.

El factor de forma rectangular (F_r) es definido mediante la siguiente ecuación:

$$F_r = \frac{A(B_r(I))}{A(I)}. \tag{11}$$

El índice de asimetría (A_i) es definido según:

$$A_i = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}|A(i,j) - F(i,j)|}{256\, m\, n} \tag{12}$$

donde A es la imagen de la galaxia, F es la imagen de la galaxia rotada 180 grados, m y n son el alto y ancho respectivamente en pixeles del mínimo rectángulo que envuelve a la galaxia. El rango de intensidad de los pixeles en escala de grises va de 0 a 255, entonces, el denominador es multiplicado por 256 para normalizar la ecuación.

Una de las principales contribuciones del presente TFG es la propuesta de otras MFs, a las cuales llamamos picos horizontales (Ph) y picos verticales (Pv) del histograma, y la firma de dispersión lumínica (FDL). También haremos uso de las propuestas realizadas en [5] Ratio de Circularidad, Ratio de Forma, Ratio de Compacidad y Ratio de Radio.

Los picos horizontales están definidos como:

$$Ph = \frac{d}{dt}Hh(Rh \cap I) \tag{13}$$

donde Rh es la recta que cruza el eje horizontal de la galaxia, I es la imagen de la galaxia y Hh es el histograma de la intersección entre Rh e I.

Los picos verticales están definidos como:

$$Pv = \frac{d}{dt} Hv(Rv \cap I) \tag{14}$$

donde Rv es la recta que cruza el eje vertical de la galaxia y Hv es el histograma de la intersección entre Rv e I.

Una desventaja que introduce (9) y es la posibilidad de variación de la proporción entre el área y el perímetro cuando se consideran formas similares de tamaños diferentes, por más que las imágenes sean de galaxias del mismo tipo. Por ejemplo (y solo a modo de analogía) si consideramos un cuadrado de lado 2 y otro de lado 3, la ecuación (9) devolvería $0{,}5$ y $0{,}75$ respectivamente. Podríamos adaptar esta medida para hacerla adimensional utilizando la siguiente ecuación:

$$Cir = \frac{A}{P^2} \tag{15}$$

donde A es el área y P el perímetro. Esto daría entonces el valor de $1/16$ para ambos cuadrados. Esta proporción adquiere su máximo valor cuando la forma del objeto es circular. En este caso usaremos (16).

$$Cir = \frac{\pi r^2}{(2\pi r)^2} = \frac{1}{4\pi} \tag{16}$$

Para hacer que la medida esté entre 0 y 1 podríamos, por tanto, escalar multiplicando por 4π. Los geógrafos se valen de esto y lo denominan ratio de circularidad (R_1):

$$R_1 = \frac{4\pi A}{P^2}. \tag{17}$$

El ratio de forma (R_2) está dado por (18)

$$R_2 = \frac{4A}{\pi l^2} \tag{18}$$

donde l es la longitud de la línea que une los dos puntos más distantes de la forma.

El ratio de compacidad (R_3) está dado por (19)

$$R_3 = \frac{A}{\pi R^2} \tag{19}$$

donde R es el radio del círculo más pequeño que rodea la forma.

El ratio de radio (R_4) está dado por (20)

$$R_4 = \frac{r}{R} \tag{20}$$

donde r es el radio del círculo mayor que pueda insertarse en la forma.

Por último, creamos el método de extracción de característica FDL, y consiste en darle un peso a los pixeles de la imagen de acuerdo a su distancia al centro de la misma. La distancia es medida en pixeles de forma horizontal o vertical a los ejes horizontal o vertical respectivamente del punto central, y la considerada es la mayor distancia entre las dos. El peso de cada pixel está dado por la siguiente ecuación:

$$FDL = \frac{1}{n} I \tag{21}$$

donde

$$n = n_{i-1} + 8_i \tag{22}$$

representa la cantidad de pixeles a una misma distancia al centro de la imagen e I es la intensidad en escala de grises del pixel considerado. De esta forma, por cada distancia al centro de la imagen, habrá un máximo de 255 de intensidad, sin importar cuantos pixeles estén a la misma distancia. El valor 8 es porque el punto central está rodeado por 8 pixeles, a la distancia de 1 pixel cada uno, y cuando aumenta la distancia, habrán 8 pixeles más que el nivel anterior. Por esto, el valor de n es igual a 8 por la distancia al eje más lejano, más el valor de n a una distancia de un pixel menos al centro de la imagen.

III-B2. Análisis de Componentes Principales: El PCA es un método estadístico que transforma un número de variables posiblemente correlacionadas a un número más pequeño de variables no correlacionadas o PCs. El PCA es usado generalmente para reducir la dimensionalidad de un conjunto de datos mientras retiene la mayor cantidad de información posible. PCA es una herramienta para buscar patrones en datos de muchas dimensiones como lo son las imágenes.

Los datos habitualmente se organizan en una matriz X de $n \times p$ dimensiones donde n es el número de filas (observaciones) y p representa el número de columnas (variables). El objetivo del PCA es sintetizar la información contenida en las p columnas de la matriz de datos X, es decir, las variables, tal que la dimensión del problema se reduzca.

El *dataset* utilizado en este TFG consiste en N imágenes de m pixeles de ancho por m pixeles de alto correspondientes a galaxias del cúmulo local utilizadas en [6]. Organizaremos nuestra matriz X en N filas u observaciones (en nuestro caso galaxias) y $m \times m$ columnas o variables (pixeles de las imágenes de galaxias remuestreadas en un vector de una sola dimensión). El PCA se lleva a cabo a través de la descomposición en valores singulares (SVD, por sus siglas en inglés) de la matriz de covarianzas $(p \times p)$. Este proceso genera la siguiente matriz de covarianza

$$C = AA^T \tag{23}$$

donde A representa una fila de la matriz X. Esta matriz C, entonces, tendrá una dimensión de $p \times p$, y sabiendo que $p = m \times m$, determinar los p^2 autovectores y autovalores se vuelve una tarea impracticable. En [7] y [8] sugieren resolver

este problema calculando la matriz de covarianza mediante la siguiente ecuación:

$$L = A^T A \tag{24}$$

ya que los autovectores de la matriz L serían una combinación lineal de los autovectores de la matriz C. *OpenCV* nos provee un método eficiente para obtener los PCs de una matriz con miles de columnas, por lo que hallaremos los PCs tanto del *dataset* original como de la traspuesta del mismo.

Un problema que podría tener el PCA es, que cuando las variables de la matriz de entrada no muestran correlación, los resultados no pueden ser muy buenos, por lo que para este caso, la técnica que se puede utilizar es una variante del PCA, el método ICA detallado a continuación.

III-B3. Análisis de Componentes Independientes: El objetivo fundamental del ICA es proporcionar un método que permita encontrar una representación lineal de los datos no gaussianos de forma que las componentes sean estadísticamente independientes o lo más independientes posible. Una representación de este tipo permite obtener la estructura fundamental de los datos en muchas aplicaciones, incluidas la extracción de características y la separación de señales.

Existen varios métodos y algoritmos que permiten obtener la matriz de transformación ICA, entre las que podemos citar *Informax, Comon's, FastICA* y *JADE* [9]. En este TFG utilizaremos el algoritmo denominado *FastICA* propuesto por *Aapo Hyvärien* en la Universidad de Tecnología de Helsinki [10]. Es un algoritmo basado en el método del punto fijo y es adecuado para su realización en lenguajes de simulación matemática. Desde el punto de vista del rendimiento de los algoritmos que implementan ICA, se ha demostrado empíricamente que existen diferencias muy pequeñas y que todos obtienen un óptimo muy similar de ICs [11].

Mientras PCA maximiza la varianza, ICA minimiza mayores órdenes de dependencia. El número de variables debe ser mayor o igual al número de observaciones, para nuestro caso, la matriz de origen debe tener mayor o igual número de columnas que de filas. Al igual que en [11], realizaremos una implementación alternativa, en la que alimentaremos al algoritmo *FastICA* con la matriz de covarianza de la traspuesta del *dataset*.

IV. Proceso de Clasificación

En esta fase, hacemos uso de las técnicas de aprendizaje automático (*machine learning*). El aprendizaje automático es una rama de la IA, cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender. Lo que hace es generalizar comportamientos a partir de una información no estructurada suministrada en forma de ejemplos, por tanto, es un proceso de inducción del conocimiento. En muchas ocasiones, este campo se solapa con el de la estadística, ya que las dos disciplinas se basan en el análisis de datos. El aprendizaje automático puede ser visto como un intento de automatizar algunas partes del método científico mediante métodos matemáticos.

Algunos sistemas de aprendizaje automático intentan eliminar toda necesidad de intuición o conocimiento experto de los procesos de análisis de datos, a estos se los denominan sistemas de aprendizaje automático no supervisados. Otros tratan de establecer un marco de colaboración entre el experto y la computadora, a estos se los denominan sistemas de aprendizaje automático supervisados.

En este TFG, utilizaremos los algoritmos de aprendizaje automático supervisado para clasificar las imágenes de galaxias disponibles en el *dataset*. Esta decisión se debe a que queremos clasificar las galaxias en base a la secuencia de *Hubble*, por lo que como expertos, tendremos que indicarle a los algoritmos de clasificación qué tipos de galaxias son las que se encuentran en el *dataset*, que será utilizado como datos de entrenamiento, y sirvan de aprendizaje para que puedan utilizar este conocimiento para poder clasificar nuevas imágenes no categorizadas.

Los resultados de los clasificadores son validados con la técnica de validación cruzada (*cross-validation*). Esta técnica es utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba. Consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones. La validación cruzada es una manera de predecir el ajuste de un modelo a un hipotético conjunto de datos de prueba cuando no disponemos del conjunto explícito de datos de prueba. Existen varios tipos de validaciones cruzadas, en este TFG, utilizaremos la validación cruzada de K iteraciones o (*K-fold cross-validation*).

Como aquí no profundizaremos en estos métodos, solo citaremos los algoritmos que utilizaremos para clasificar con la herramienta WEKA [12]. En la **Figura 8** vemos cómo están agrupados los algoritmos utilizados para clasificación en este TFG.

V. Resultados Experimentales

En este capítulo mostramos los resultados de aplicar clasificación con los distintos tipos de características estudiadas en este TFG. Una vez extraídas todas las MFs, PCs e ICs, procedemos a comparar los resultados. Para realizar la clasificación, agrupamos las galaxias en grupos de tres (S, E, Irr), cinco (E, S0, Sa+Sb, Sc+Sd, Irr) y siete (E, S0, Sa, Sb, Sc, Sd, Irr) clases, y evaluamos los distintos algoritmos de clasificación.

V-A. Clasificación Usando Características Morfológicas

Para comparar el desempeño de los algoritmos de clasificación con las MFs, probamos con cada característica

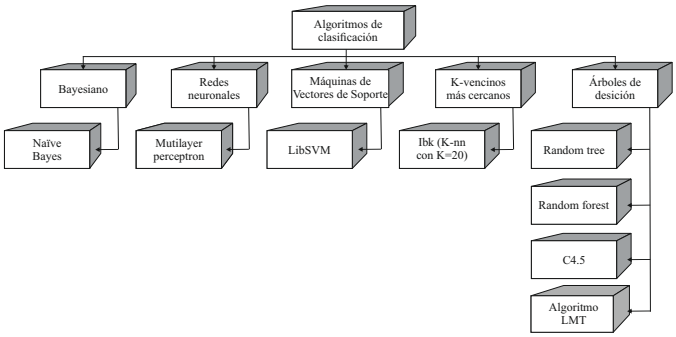

Figura 8: Algoritmos de clasificación agrupados por tipo

por separado, y luego realizamos una selección de características cuya combinación nos daría mejores resultados que utilizándolas por separado. Para la selección de características utilizamos el método de selección de atributos de WEKA realizando búsqueda exhaustiva (*ExhaustiveSearch*) y como evaluador la selección de atributos basada en correlación (CFS, por sus siglas en inglés). Con esto, realizamos 2^n combinaciones de características, donde n indica el número de características disponibles, para así buscar la mejor combinación de las mismas, es decir, las que proveen el mejor porcentaje de aciertos. En la **Tabla I** podemos ver los resultados de realizar clasificación con los atributos por separado y en la **Tabla II** vemos los resultados de realizar selección de atributos con búsqueda exhaustiva. Podemos notar, que algunos atributos dan buenos porcentajes de aciertos con ciertos algoritmos de clasificación, y con otros algoritmos dan porcentajes de aciertos más bajos. También podemos ver que usando atributos combinados, podemos obtener un mejor porcentaje de aciertos que utilizando atributos de forma individual.

En la **Tabla I** el porcentaje de aciertos más alto alcanzado, fue con la característica IA y el algoritmo de clasificación RF. En promedio, la características con mejores resultados es el IA, y los algoritmos con mejores resultados son SVM, MP, K-nn y LMT.

En la **Tabla II** podemos ver que el máximo porcentaje de aciertos se logró combinando IA, PH y RCirc, junto con el algoritmo de clasificación MP alcanzando un máximo de $86{,}58\,\%$ de efectividad. Este tipo de búsqueda de combinaciones solo se puede realizar cuando la cantidad de atributos es relativamente pequeña, porque a medida que vamos agregando más atributos, el tiempo de ejecución se duplica por cada nueva característica. Combinar atributos para obtener mejores resultados tampoco es una regla, ya que podemos notar que algunos algoritmos devuelven porcentajes inferiores que utilizando atributos individuales. Por ejemplo, con los algoritmos C4.5 y RF obtenemos porcentajes más bajos combinando atributos que con atributos por separado como vemos en la **Tabla II**.

	3-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
El	80.48	80.48	80.48	80.48	80.48	80.48	80.48
FF	80.48	80.48	80.48	80.48	80.48	80.48	80.48
Conv	80.48	73.17	80.48	80.48	80.48	79.26	80.48
FFR	80.48	80.48	80.48	80.48	80.48	80.48	80.48
IA	76.82	**81.70**	80.48	**82.92**	80.48	**82.92**	**82.92**
PH	79.26	78.04	**81.70**	80.48	**81.70**	78.04	**82.92**
PV	80.48	80.48	76.82	80.48	80.48	80.48	80.48
RCir	80.48	70.73	80.48	80.48	**81.70**	80.48	80.48
RF	80.48	80.48	80.48	80.48	80.48	80.48	80.48
RCom	80.48	79.26	80.48	79.26	80.48	76.82	**81.70**
RR	78.04	78.04	80.48	80.48	80.48	78.04	78.04
FDL	80.48	80.48	80.48	80.48	80.48	80.48	80.48

Tabla I: Clasificación de galaxias S, E e Irr con atributos por separado

	3-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
IA PH RCir	81.70	**85.36**	84.14	**86.58**	84.14	79.26	**85.36**

Tabla II: Clasificación de galaxias S, E e Irr combinando los mejores atributos

A medida que aumentamos el número de tipos de galaxias que queremos clasificar, van disminuyendo nuestros porcentajes de aciertos. En la **Tabla III**, vemos que el mejor porcentaje de aciertos se logró con IA y K-nn. En promedio, el atributo con mejores resultados fue IA, y el algoritmo con mejores resultados fue BN. En la **Tabla IV** podemos ver que con la combinación de los mejores atributos, no se logró un resultado mejor que el máximo de atributos individuales, aunque, en promedio, los atributos combinados tienen mejores porcentajes de aciertos.

	5-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
El	43.90	30.48	43.90	43.90	40.24	41.46	43.90
FF	43.90	45.12	40.24	45.12	42.68	41.46	43.90
Conv	43.90	42.68	43.90	36.58	36.58	43.90	32.92
FFR	43.90	42.68	43.90	45.12	41.46	43.90	41.46
IA	**53.65**	46.34	43.90	**52.43**	**57.31**	48.78	50.00
PH	42.68	47.56	47.56	46.34	43.90	41.46	**52.43**
PV	43.90	32.92	45.12	45.12	41.46	47.56	45.12
RCir	46.34	42.68	43.90	36.58	36.58	42.68	39.02
RF	43.90	31.70	37.80	40.24	40.24	41.46	43.90
RCom	43.90	43.90	43.90	45.12	43.90	48.78	43.90
RR	43.90	42.68	43.90	46.34	45.12	47.56	46.34
FDL	43.90	50.00	46.34	48.78	46.34	37.80	40.24

Tabla III: Clasificación de galaxias E, S0, Sa+Sb, Sc+Sd e Irr con atributos por separado

	5-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
IA RCir	**53.65**	41.46	45.12	47.56	**51.21**	**52.43**	47.56

Tabla IV: Clasificación de galaxias E, S0, Sa+Sb, Sc+Sd e Irr combinando los mejores atributos

En la **Tabla V** podemos ver que el máximo porcentaje de aciertos se logró con las características IA y PH juntos con el algoritmo MP. En promedio, el atributo con mejores aciertos es IA y el algoritmo con mejor desempeño es el MP. En la **Tabla VI** vemos que la combinación de los mejores atributos no supera al máximo logrado con los atributos por separado.

	7-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
El	24.39	13.41	21.95	26.82	15.85	25.60	24.39
FF	24.39	19.51	26.82	26.82	28.04	31.70	31.70
Conv	24.39	25.60	21.95	32.92	26.82	30.48	26.82
FFR	24.39	19.51	23.17	21.95	28.04	19.51	19.51
IA	30.48	**35.36**	26.82	**42.68**	**39.02**	**34.14**	**39.02**
PH	24.39	31.70	30.48	**42.68**	**39.02**	28.04	29.26
PV	24.39	30.48	21.95	30.48	32.92	24.39	32.92
RCir	25.60	21.95	23.17	**36.58**	**36.58**	31.70	29.26
RF	24.39	26.82	18.29	24.39	20.73	20.73	23.17
RCom	24.39	26.82	25.60	30.48	29.26	26.82	29.26
RR	24.39	32.92	26.82	31.70	31.70	23.17	25.60
FDL	24.39	26.82	29.26	24.39	30.48	29.26	23.17

Tabla V: Clasificación de galaxias E, S0, Sa, Sb, Sc, Sd e Irr con atributos por separado

	7-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
IA RCir	31.70	29.26	24.39	**34.14**	30.48	30.48	**37.80**

Tabla VI: Clasificación de galaxias E, S0, Sa, Sb, Sc, Sd e Irr combinando los mejores atributos

V-B. Clasificación Usando Componentes Principales

Para comparar el desempeño de los algoritmos de clasificación con los PCs, probamos con cada PC por separado, y luego realizamos una selección de PCs cuya combinación nos daría mejores resultados que utilizándolos por separado. Para la selección de PCs utilizamos el método de selección de atributos de WEKA realizando búsqueda con el algoritmo "*GreedyStepwise*", ya que no podemos realizar una búsqueda exhaustiva debido al tiempo que tomaría evaluar las 2^{82} combinaciones posibles. Por esto, tal vez los mejores resultados alcanzados con este algoritmo de búsqueda no sean los mejores que se podrían alcanzar con una búsqueda exhaustiva.

Existen tantos PCs como imágenes en el *dataset*, pero los PCs calculados están ordenados por su relevancia, esto es, por el porcentaje de variabilidad que determinan sus autovalores. Como los primeros cinco componentes determinan el $93{,}97\,\%$ de la variabilidad de los datos, solo utilizamos los cinco primeros componentes para realizar las pruebas individuales, luego realizamos una evaluación considerando los 82 PCs y también realizamos selección de atributos.

Como utilizamos dos métodos para extracción de PCs, en las **Tablas VII** al **XV** mostramos los resultados de utilizar los PCs del *dataset* y en las **Tablas XVI** al **XXIV** mostramos los resultados de utilizar los PCs de la traspuesta del *dataset*. Las **Tablas VII, VIII** y **IX** muestran los resultados de clasificar las galaxias considerando tres tipos, S, E e Irr.

	3-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
PC1	80.49	80.49	80.49	80.49	80.49	80.49	80.49
PC2	80.49	80.49	80.49	80.49	80.49	80.49	80.49
PC3	80.49	79.27	80.49	80.49	80.49	80.49	80.49
PC4	80.49	78.05	80.49	80.49	80.49	80.49	80.49
PC5	80.49	80.49	79.27	80.49	80.49	80.49	80.49

Tabla VII: Clasificación de galaxias E, S e Irr con los PCs por separado

En la **Tabla VII** vemos que la mayoría de los algoritmos de clasificación tienen una efectividad similar, alcanzando aciertos de $80{,}49\,\%$, pero los algoritmos RF y RT tienen rendimientos más bajos. En la **Tabla VIII** podemos ver que el algoritmo NB destaca sobre los demás con un nivel de acierto del $82{,}92\,\%$, y los algoritmos que mostraban un rendimiento bajo cuando consideramos los PCs por separado, RF y RT, mejoran su rendimiento al considerar todos los PCs.

	3-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
Todos	78.04	**82.92**	80.48	80.48	80.48	76.82	74.39

Tabla VIII: Clasificación de galaxias E, S e Irr usando todos los PCs

En la **Tabla IX** vemos los resultados de la clasificación obtenidos con los PCs seleccionados utilizando selección de atributos. Podemos notar que el algoritmo NB mantiene una pequeña ventaja sobre los demás y sólo NB, SVM y K-nn obtienen resultados superiores a $80\,\%$. Los componentes seleccionados dependen del algoritmo *GreedyStepwise* utilizado para realizar la búsqueda de los mejores atributos.

	3-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
PC15 PC75 PC77 PC79	78.04	**82.92**	80.48	78.04	80.48	76.82	78.04

Tabla IX: Clasificación de galaxias E, S e Irr seleccionando los mejores PCs

En las **Tablas X, XI** y **XII** vemos los resultados de clasificar las galaxias considerando cinco tipos, E, S0, Sa+Sb, Sc+Sd e Irr. En la **Tabla X** vemos los resultados de clasificar considerando los PCs por separado, el mejor resultado se alcanzó con el algoritmo K-nn, sin embargo, en promedio no fue el mejor, ya que quienes alcanzaron el promedio más alto fueron los algoritmos BN y LMT. También podemos ver que los algoritmos RF y RT tienen muy bajo rendimiento.

	5-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
PC1	47.56	45.12	43.90	43.90	43.90	46.34	50.00
PC2	43.90	36.59	42.68	39.02	36.59	37.80	40.24
PC3	43.90	43.90	41.46	45.12	**51.22**	48.78	47.56
PC4	43.90	40.24	43.90	41.46	37.80	41.46	42.68
PC5	43.90	42.68	43.90	41.46	35.37	45.12	40.24

Tabla X: Clasificación de galaxias E, S0, Sa+Sb, Sc+Sd e Irr con los PCs por separado

En la **Tabla XI** vemos el resultado utilizando todos los PCs, y podemos notar que al igual que para la clasificación de tres tipos, el algoritmo con mejor resultado es el NB alcanzando $56{,}1\,\%$ y los algoritmos RF y RT mejoraron sus resultados.

	5-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
Todos	46.34	**56.10**	43.90	43.90	43.90	42.68	41.46

Tabla XI: Clasificación de galaxias E, S0, Sa+Sb, Sc+Sd e Irr usando todos los PCs

En la **Tabla XII** vemos los resultados luego de aplicar selección de atributos. En este caso, es el algoritmo RF el que obtiene el mejor resultado, alcanzando $56{,}1\,\%$ de aciertos, igualando al resultado del NB con todos los PCs.

	5-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
PC1 PC15 PC61 PC72 PC75 PC79 PC80	46.34	**51.22**	43.90	43.90	37.80	50.00	46.34

Tabla XII: Clasificación de galaxias E, S0, Sa+Sb, Sc+Sd e Irr seleccionando los mejores PCs

En las **Tablas XIII, XIV** y **XV** vemos los resultados de clasificar las galaxias considerando siete tipos, E, S0, Sa, Sb, Sc, Sd e Irr. En la **Tabla XIII** vemos los resultados de clasificar considerando los PCs por separado, los porcentajes de aciertos más altos se alcanzaron con los algoritmos NB, MP y K-nn, pero el algoritmo con mejor promedio es el NB.

	7-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
PC1	23.17	31.71	24.39	26.83	28.05	20.73	30.49
PC2	24.39	28.05	29.27	26.83	20.73	26.83	28.05
PC3	24.39	**34.15**	23.17	32.93	32.93	23.17	29.27
PC4	24.39	30.49	25.61	**34.15**	**34.15**	21.95	19.51
PC5	23.17	29.27	29.27	24.39	25.61	25.61	23.17

Tabla XIII: Clasificación de galaxias E, S0, Sa, Sb, Sc, Sd e Irr con los PCs por separado

En la **Tabla XIV** vemos los resultados de clasificar considerando todos los PCs, el porcentaje más alto se logró con el algoritmo NB, seguido por el RF. El resto de algoritmos dieron resultados muy bajos.

	7-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
Todos	23.17	**43.90**	24.39	26.83	24.39	29.27	26.83

Tabla XIV: Clasificación de galaxias E, S0, Sa, Sb, Sc, Sd e Irr usando todos los PCs

En la **Tabla XV** vemos el resultado de clasificar utilizando selección de atributos, el algoritmo con mejor resultado es el NB pero el porcentaje de aciertos es muy inferior que considerando todos los PCs.

	7-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
PC35 PC80	23.17	**34.15**	24.39	21.95	28.05	21.95	20.73

Tabla XV: Clasificación de galaxias E, S0, Sa, Sb, Sc, Sd e Irr seleccionando los mejores PCs

Los PCs no tienen el mismo comportamiento en un sistema de clasificación como lo tienen en un sistema de reconstrucción de imágenes. El PCA se hace con la intención de reducir la dimensionalidad, pero como podemos ver en estas Tablas, usar el segundo, tercero, cuarto o quinto componente principal, no tiene casi diferencias con respecto al primero. Usar todos los PCs disponibles, o los atributos seleccionados podrían mejorar los resultados, pero tampoco significa que la combinación de atributos sea siempre mejor que un atributo individual. No podemos determinar de forma anticipada cuál sería la combinación perfecta de PCs que nos ayude a obtener los mejores resultados. Pero si utilizamos un *dataset* con mayor número de observaciones, los mejores resultados serán con los PCs con mayor varianza.

En las **Tablas XVI** al **XXIV** vemos los resultados de aplicar PCA a la traspuesta del *dataset*. Como las pruebas anteriores, el comportamiento es el mismo, devuelve buenos resultados al clasificar pocos tipos de galaxias pero va disminuyendo su efectividad a medida que aumentamos los tipos de galaxias que queremos clasificar. Las **Tablas XVI, XVII** y **XVIII** muestran los resultados de clasificar las galaxias considerando tres tipos, S, E e Irr. En la **Tabla XVI** se consideran los PCs por separado, y como se puede ver, los resultados son casi iguales a los de la **Tabla VII**, siendo igualmente los algoritmos RF y RT los de resultados más bajos.

	3-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
PC1	80.49	80.49	80.49	80.49	80.49	80.49	80.49
PC2	80.49	80.49	80.49	80.49	80.49	79.27	80.49
PC3	80.49	80.49	80.49	79.27	80.49	79.27	80.49
PC4	80.49	79.27	80.49	80.49	80.49	80.49	80.49
PC5	80.49	79.27	80.49	80.49	80.49	76.83	80.49

Tabla XVI: Clasificación de galaxias E, S e Irr con los PCs por separado

La **Tabla XVII** sí muestra una mejora respecto a la **Tabla VIII**, se puede ver que el resultado del algoritmo NB es $6{,}28\,\%$ mejor, alcanzando un $89{,}02\,\%$ de aciertos. También podemos ver que los algoritmos que tienen un bajo rendimiento cuando consideramos los PCs por separado, como RF y RT, tienen mejores resultados al considerar todos los PCs.

	3-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
Todos	73.17	**89.02**	80.49	**81.71**	80.49	75.61	70.73

Tabla XVII: Clasificación de galaxias E, S e Irr usando todos los PCs

En la **Tabla XVIII** vemos los resultados de luego de aplicar selección de atributos. Los algoritmos con mejores resultados fueron SVM y K-nn alcanzando $80{,}49\,\%$ de efectividad.

	3-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
PC3 PC11 PC15 PC42 PC76 PC78 PC80	73.17	79.27	80.49	76.83	80.49	73.17	76.83

Tabla XVIII: Clasificación de galaxias E, S e Irr seleccionando los mejores PCs

Las **Tablas XIX, XX** y **XXI** muestran los resultados de clasificar las galaxias considerando cinco tipos, E, S0, Sa+Sb, Sc+Sd e Irr. En la **Tabla XIX** vemos el resultado de la clasificación considerando los PCs por separado. Podemos notar que el mejor porcentaje de aciertos se logró con el algoritmo K-nn, sin embargo, los algoritmos con mejores promedios son NB y LMT.

	5-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
PC1	37.80	45.12	43.90	41.46	48.78	40.24	48.78
PC2	43.90	43.90	43.90	41.46	41.46	37.80	40.24
PC3	37.80	47.56	43.90	43.90	48.78	34.15	48.78
PC4	43.90	41.46	43.90	39.02	34.15	42.68	42.68
PC5	43.90	43.90	43.90	37.80	42.68	37.80	40.24

Tabla XIX: Clasificación de galaxias E, S0, Sa+Sb, Sc+Sd e Irr con los PCs por separado

En la **Tabla XX** vemos el resultado de la clasificación considerando todos los PCs, y los mejores resultados logrados fueron con los algoritmos NB y RF, alcanzando $57{,}32\,\%$ de efectividad.

	5-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
Todos	43.90	**57.32**	43.90	43.90	43.90	**53.66**	41.46

Tabla XX: Clasificación de galaxias E, S0, Sa+Sb, Sc+Sd e Irr usando todos los PCs

En la **Tabla XXI** vemos los resultados de la clasificación luego de realizar selección de atributos. En este caso, el mejor porcentaje de aciertos se logró con el NB alcanzando $59{,}76\,\%$ de efectividad. Pero si comparamos con la **Tabla XII**, podemos ver que no siempre la selección de atributos nos devolverá un resultado mejor que considerando todos los PCs.

	5-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
PC1 PC3 PC62 PC73 PC76 PC78 PC80 PC81	43.90	**59.76**	43.90	40.24	50.00	42.68	47.56

Tabla XXI: Clasificación de galaxias E, S0, Sa+Sb, Sc+Sd e Irr seleccionando los mejores PCs

Las **Tablas XXII, XXIII** y **XXIV** muestran los resultados de clasificar las galaxias considerando siete tipos, E, S0, Sa, Sb, Sc, Sd e Irr. En la **Tabla XXII** vemos los resultados de la clasificación considerando los PCs por separado, tanto el mejor resultado como el mejor promedio se alcanzó con el algoritmo K-nn con un máximo de $40{,}24\,\%$ de aciertos.

	7-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
PC1	24.39	29.27	24.39	30.49	26.83	17.07	30.49
PC2	24.39	25.61	24.39	21.95	24.39	18.29	26.83
PC3	24.39	31.71	24.39	31.71	28.05	21.95	24.39
PC4	24.39	31.71	24.39	**37.80**	32.93	30.49	**35.37**
PC5	24.39	15.85	24.39	21.95	**40.24**	26.83	23.17

Tabla XXII: Clasificación de galaxias E, S0, Sa, Sb, Sc, Sd e Irr con los PCs por separado

En la **Tabla XXIII** vemos los resultados de clasificar considerando todos los PCs. El algoritmo con mejor resultado es el NB con $41{,}46\,\%$ de aciertos seguido por LMT con $39{,}02\,\%$ de aciertos.

	7-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
Todos	30.49	**41.46**	24.39	**34.15**	24.39	32.93	**39.02**

Tabla XXIII: Clasificación de galaxias E, S0, Sa, Sb, Sc, Sd e Irr usando todos los PCs

En la **Tabla XXIV** podemos notar algo muy particular, y es que la selección de atributos devolvió un solo atributo como mejor combinación, dando al NB el mejor resultado.

	7-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
PC81	30.49	**37.80**	24.39	28.05	30.49	25.61	**34.15**

Tabla XXIV: Clasificación de galaxias E, S0, Sa, Sb, Sc, Sd e Irr seleccionando los mejores PCs

Podemos notar que si consideramos la traspuesta de la matriz del *dataset*, tenemos mejores resultados al clasificar tres y cinto tipos de galaxias que si consideramos el *dataset* original, además, el cálculo de los PCs es mucho más simple porque la matriz de covarianza solo tiene 82 filas por 82 columnas, pero al clasificar siete tipos de galaxias, tuvimos resultados más bajos que considerando el *dataset* original. También podemos notar que los algoritmos RF y RT mejoran sustancialmente cuando se consideran todos los PCs. Algo llamativo es que el algoritmo SVM mantiene casi el mismo porcentaje de aciertos tanto para PCs por separado, todos los PCs juntos o con PCs obtenidos con selección de atributos. Por más que los resultados del SVM no sean los mejores, son buenos resultados y su constancia sería una ventaja para utilizarlo con PCs.

V-C. Clasificación Usando Componentes Independientes

En las **Tablas XXV** al **XXXIII** vemos los resultados de utilizar ICs para realizar la clasificación. Las **Tablas XXV, XXVI** y **XXVII** muestran los resultados de clasificar las galaxias considerando tres tipos, S, E e Irr. En la **Tabla XXV** vemos los resultados de clasificar considerando los ICs por separado. Los porcentajes son similares a los obtenidos utilizando PCs, por lo que no se puede apreciar ninguna ventaja o mejora.

	3-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
IC1	80.49	78.05	80.49	80.49	80.49	80.49	80.49
IC2	80.49	80.49	80.49	80.49	80.49	80.49	80.49
IC3	80.49	79.27	80.49	80.49	80.49	80.49	80.49
IC4	80.49	79.27	80.49	80.49	80.49	80.49	80.49
IC5	80.49	80.49	80.49	76.83	80.49	74.39	79.27

Tabla XXV: Clasificación de galaxias E, S e Irr con los ICs por separado

En la **Tabla XXVI** vemos los resultados de clasificar considerando todos los ICs, algo notable es que el algoritmo NB que suele dar buenos resultados con PCs, tiene muy bajo porcentaje de aciertos con ICs. En este caso, los algoritmos MP y RF son los que devuelven mejores porcentajes de aciertos, alcanzando $82{,}93\,\%$ de aciertos.

	3-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
Todos	71.95	65.85	80.49	**82.93**	80.49	80.49	80.49

Tabla XXVI: Clasificación de galaxias E, S e Irr usando todos los ICs

En la **Tabla XXVII** vemos los resultados de la clasificación luego de realizar selección de atributos. En esta ocasión, podemos ver que el mejor resultado se obtuvo con el algoritmo C4.5.

	3-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
IC2 IC18 IC34 IC66 IC76 IC80	79.27	60.98	80.49	78.05	80.49	81.71	78.05

Tabla XXVII: Clasificación de galaxias E, S e Irr seleccionando los mejores ICs

Las **Tablas XXVIII, XXIX** y **XXX** muestran los resultados de clasificar las galaxias considerando cinco tipos, E, S0, Sa+Sb, Sc+Sd e Irr. En la **Tabla XXVIII** vemos los resultados de la clasificación considerando los ICs por separado. Los mejores resultados se obtuvieron con los algoritmos MP y K-nn, alcanzando $54{,}88\,\%$ de aciertos. El algoritmo con mejor promedio es el K-nn. De nuevo, los algoritmos RF y RT, al igual que con PCs, tienen muy baja efectividad.

	5-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
IC1	45.12	48.78	50.00	41.46	**54.88**	41.46	46.34
IC2	43.90	43.90	42.68	36.59	40.24	39.02	42.68
IC3	43.90	**51.22**	40.24	43.90	46.34	42.68	43.90
IC4	43.90	34.15	43.90	41.46	37.80	41.46	37.80
IC5	40.24	**52.44**	47.56	**54.88**	**53.66**	47.56	**53.66**

Tabla XXVIII: Clasificación de galaxias E, S0, Sa+Sb, Sc+Sd e Irr con los ICs por separado

En la **Tabla XXIX** vemos los resultados de la clasificación considerando todos los ICs. El mejor resultado se alcanza con el algoritmo SVM, llegando a $54{,}88\,\%$ de aciertos.

	5-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
Todos	47.56	45.12	**54.88**	47.56	43.90	50.00	**52.44**

Tabla XXIX: Clasificación de galaxias E, S0, Sa+Sb, Sc+Sd e Irr usando todos los ICs

En la **Tabla XXXI** vemos los resultados luego de aplicar selección de atributos, en este caso, el mejor porcentaje de aciertos se logró con el algoritmo SVM alcanzando $52{,}44\,\%$.

	5-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
IC18 IC34 IC68 IC80	43.90	45.12	**52.44**	47.56	47.56	46.34	45.12

Tabla XXX: Clasificación de galaxias E, S0, Sa+Sb, Sc+Sd e Irr seleccionando los mejores ICs

Las **Tablas XXXI, XXXII** y **XXXIII** muestran los resultados de clasificar las galaxias considerando siete tipos, E, S0, Sa, Sb, Sc, Sd e Irr. En la **Tabla XXXI** vemos los resultados de la clasificación considerando los ICs por separado, dando mejores resultados con los algoritmos NB y K-nn, alcanzando $40{,}24\,\%$ de aciertos.

	7-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
IC1	21.95	**35.37**	**34.15**	30.49	**35.37**	24.39	**36.59**
IC2	24.39	**36.59**	28.05	32.93	26.83	30.49	25.61
IC3	24.39	**40.24**	25.61	**36.59**	**40.24**	30.49	29.27
IC4	24.39	28.05	30.49	30.49	29.27	15.85	19.51
IC5	23.17	31.71	**34.15**	30.49	32.93	29.27	32.93

Tabla XXXI: Clasificación de galaxias E, S0, Sa, Sb, Sc, Sd e Irr con los ICs por separado

En la **Tabla XXXII** vemor los resultados de la clasificación considerando todos los ICs. El mejor resultado se logró con el algoritmo LMT llegando a alcanzar $39{,}02\,\%$ de aciertos.

	7-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
Todos	23.17	32.93	**34.15**	32.93	24.39	30.49	**39.02**

Tabla XXXII: Clasificación de galaxias E, S0, Sa, Sb, Sc, Sd e Irr usando todos los ICs

En la **Tabla XXXIII** vemos que hay un solo IC luego de la selección de atributos, dando el mejor resultado con el algoritmo LMT, con $35{,}37\,\%$ de aciertos.

	7-class						
	BN	NB	SVM	MP	K-nn	C4.5	LMT
IC80	23.17	**34.15**	29.27	32.93	29.27	31.71	**35.37**

Tabla XXXIII: Clasificación de galaxias E, S0, Sa, Sb, Sc, Sd e Irr seleccionando los mejores ICs

Los ICs tienen el mismo comportamiento que los PCs para la clasificación, con la diferencia que la calidad de los resultados son de menor calidad.

VI. Conclusiones

Este Trabajo de Fin de Grado ha abordado el estudio de la clasificación automática de galaxias. Las principales contribuciones y conclusiones se resumen a continuación:

- La clasificación de galaxias es un tema actual de estudio, las técnicas de *machine learning* han demostrado ser efectivas para tareas donde se manejan grandes volúmenes de datos. La calidad de los resultados siempre dependerá del previo preprocesamiento de imágenes y extracción de características, como así también la selección de buenos datos de entrenamiento.
- Las MFs más efectivas resultaron ser aquellas que tienen en cuenta el contenido de la imagen, como IA y PH, y no solo la forma de los mismos, como El, Conv y el RForm. La propuesta realizada FDL es una característica que tiene en cuenta el interior de la imagen, pero no resultó ser mejor que IA o PH, esto podría ser porque solo tiene en cuenta la distancia de los pixeles al centro de la imagen y no su orientación.
- Los PCs con mayor varianza dan resultados más acertados que los PCs con menor varianza, aunque en estas pruebas no pudimos notar tanta diferencia debido al tamaño de nuestro *dataset* y también porque los datos están desbalanceados. La cantidad de PCs que se deben considerar dependen de los autovalores, donde un autovalor mayor a 1 indica que el PC representa mayor varianza que la contabilidaza por una de las variables originales. Algo que podemos resaltar es la mejora en la calidad de los resultados si utilizamos la traspuesta del *dataset*, y esto es algo llamativo, ya que los PCs son los autovectores de la matriz de covarianza, por lo que en este caso, la matriz de covarianza L sería una matriz de dimensiones 82 x 82 y procesar esta matriz es bastante rápido.
- Los ICs muestran el mismo comportamiento que los PCs, solo que con una calidad de resultados un poco inferior.
- Los resultados de estas pruebas están sujetas a los datos de entrenamiento, y los porcentajes de aciertos podrían ser un poco sensibles debido a la cantidad de tipos de galaxias que utilizamos, por lo que para tener una mejor evaluación, los datos de entrenamiento deberían ser más equilibrados, es decir, poseer un *dataset* con cantidades equivalentes de galaxias espirales, elípticas e irregulares. También sería conveniente utilizar más imágenes de entrenamiento en el *dataset*.
- Entre las MFs, los PCs y los ICs, son las MFs las que nos aseguran una buena clasificación. La cuestión sería utilizar una MF o una combinación de características que arrojen buenos resultados con los datos de entrenamientos, y para esto, deberíamos escoger características que tengan en cuenta la superficie de la imagen.
- Una buena característica es aquella que separa correctamente a las observaciones de una muestra, por ejemplo, si tenemos un grupo de 10 personas adultas y 10 niños, la altura sería una buena característica ya que sirve para diferenciar los dos grupos, de esta forma, dada la altura de un individuo, con dicho dato se podría clasificar fácilmente a qué grupo o clase pertenece. Para el caso de las galaxias, es necesario una característica que diferencie bien entre espirales, elípticas e irregulares, y las características estudiadas en este TFG, aunque realizan una buena separación entre clases, existen muchos valores atípicos (*outliers*) que hacen que los clasificadores

confundan los tipos.

- Si sumamos los promedios de los algoritmos de clasificación, el *Naïve Bayes* es el que provee los mejores resultados en comparación a los demás métodos, con 10 promedios superiores, seguido por SVM, K-nn, LMT y RF con 3 promedios superiores cada uno, BN y C4.5 solo con 1 promedio superior cada uno. El algoritmo RT no fue mejor en ningún caso. Algo que también podemos notar es que el método SVM es más efectivo con ICs.
- Los algoritmos RF y RT muestran bajos niveles de efectividad cuando se consideran atributos por separado, pero si se considera una combinación de MFs o muchos PCs o ICs mejoran sustancialmente. El algoritmo SVM muestra un comportamiento muy estable, devolviendo buenos valores para la mayoría de los casos.

REFERENCIAS

[1] Eggen, O. J.; Lynden-Bell, D.; Sandage, A. R., "Evidence from the motions of old stars that the Galaxy collapsed", Astrophysical Journal, vol. 136, p. 748, November, 1962.

[2] S. Kasivajhula, N. Raghavan, H. Shah. "Morphological Galaxy Classification Using Machine Learning". 2007.

[3] J. de la Calleja, O. Fuentes. "Automated Classification of Galaxy Images," Instituto Nacional de Astrofísica, Óptica y Electrónica, Luís Enrique Erro 1. 2004.

[4] International Telecommunication Union, "Recommendation BT. 601", Available from: https://www.itu.int. Retrieved August, 2016.

[5] K. E. Selkirk, "Pattern and Place: An Introduction to the Mathematics of Geography", Cambridge Univ. Press, Cambridge, New York, 1982.

[6] Z. Frei, "Automatic morphological classification of galaxies" Institute of Physics, Eötvös University. 1999.

[7] M. A. Turk, "Face Recognition Using Eigenfaces" Vision and Modeling Group, The Media Laboratory, Massachusetts Institute of Technology. 1991.

[8] M. A. Turk, Alex Pentland, "Eigenfaces for Recognition" Vision and Modeling Group, The Media Laboratory, Massachusetts Institute of Technology. 1991.

[9] M. A. Vicente, C. Fernández, A. Gil, L. Pavá, "Equivalencia entre ICA y PCA como métodos de extracción de características en reconocimiento visual basado en apariencia", Dpto. de Ingeniería de Sistemas Industriales, Universidad Miguel Hernández. Alicante. España. 2007

[10] Aapo Hyvärinen, Juha Karhunen, Erkki,Oja. "Independent component analysis" (1st ed.). New York: J. Wiley, 2001.

[11] M. Delbracio, M. Mateu, "Identificación utilizando PCA, ICA y LDA". 2006. Lior Rokach, Oded Maimon. "Data mining with decision trees: theory and applications", World Scientific, 2008.

[12] Witten, I. H., Frank, E., "Data mining: practical machine learning tools with Java implementations", Morgan Kaufmann, San Francisco, 2000.