Universidad Autónoma del Estado de Hidalgo

Instituto de Ciencias Básicas e Ingeniería

licenciatura en Ingeniería en Telecomunicaciones

Clasificación automática de galaxias empleando imágenes digitales del S.D.S.S.

T E S I S

que para obtener el título de:

Licenciado en Ingeniería en telecomunicaciones

presenta:

Cortés Ávila Jorge Alfredo

Directores:

Dr. Omar López Ortega
Dr. Pedro Amado Miranda Romagnoli

Mineral de la Reforma, Hidalgo, 2024

Universidad Autónoma del Estado de Hidalgo
Instituto de Ciencias Básicas e Ingeniería
School of Engineering and Basic Sciences

Mineral de la Reforma, Hgo., a 6 de marzo de 2024

Número de control: ICBI-D/341/2024
Asunto: Autorización de impresión.

MTRA. OJUKY DEL ROCÍO ISLAS MALDONADO
DIRECTORA DE ADMINISTRACIÓN ESCOLAR DE LA UAEH

Con fundamento en lo dispuesto en el Título Tercero, Capítulo I, Artículo 18 Fracción IV; Título Quinto, Capítulo II, Capítulo V, Artículo 51 Fracción IX del Estatuto General de nuestra Institución, por este medio le comunico que el Jurado asignado al Pasante de la Licenciatura en Ingeniería en Telecomunicaciones **Jorge Alfredo Cortes Ávila,** quien presenta el trabajo de titulación **Clasificación automática de galaxias empleando imágenes digitales del S. D.S. S.",** después de revisar el trabajo en reunión de Sinodales ha decidido autorizar la impresión del mismo, hechas las correcciones que fueron acordadas.

A continuación, firman de conformidad los integrantes del Jurado:

Presidente Dr. José Luis González Vidal

Secretario: Dr. Pedro Miranda Romagnoli

Vocal: Dr. Omar López Ortega

Suplente: Dr. Félix Agustín Castro Espinoza

Sin otro particular por el momento, reciba un cordial saludo.

Atentamente
"Amor, Orden y Progreso"

Dr. Otilio Arturo Acevedo Sandoval
Director del ICBI

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE HIDALGO
ICBI
DIRECCIÓN

OAAS/YCC

Ciudad del Conocimiento, Carretera Pachuca-Tulancingo Km. 4.5 Colonia Carboneras, Mineral de la Reforma, Hidalgo, México. C.P. 42184
Teléfono: 771 71 720 00 Ext. 2231 Fax 2109
direccion_icbi@uaeh.edu.mx

Dedicatoria ...

El resultado de este trabajo se lo dedico a mi familia. Alicia Avila, Bernardino, Alicia Islas, Antonio, Kareny y Gael.

Agradecimientos

A todas esas personas brillantes...

A mi familia, que fueron pieza clave en este proceso, por el apoyo incondicional en todo el trayecto. A mis amigos y compañeros de buenos momentos. A mis maestros de licenciatura, de todos me llevo algo.

A mis directores de tesis, por guiarme en este gran camino, por la paciencia de cada lunes, por todos los comentarios y correcciones.

En general, a todas las personas que dedican las mejores horas de su día al desarrollo de la ciencia y tecnología.

Índice general

Índice de figuras

Introducción

En el año 2019, la NASA publicó una imagen sin precedentes. El telescopio espacial Hubble capturó el débil resplandor de aproximadamente 265,000 galaxias, algunas de ellas nunca antes vistas, muchas de las galaxias están tan lejos que su luz ha tardado miles de millones de años en llegar hasta nosotros. La Fig 1. Que se presenta a continuación, se trata de casi 7500 exposiciones, lo que representa 16 años de observaciones, donde cada punto representa una galaxia.

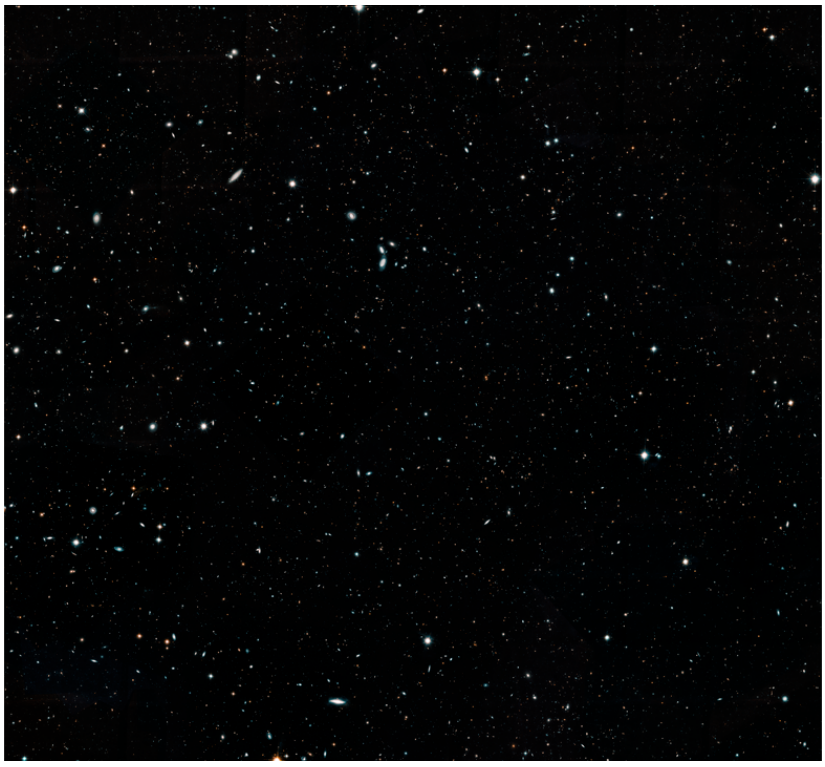


Figura 1: Hubble Legacy Field. Créditos: NASA, ESA, G. Illingworth and D. Magee, K. Whitaker, R. Bouwens, P. Oesch and the Hubble Legacy Field team.

El cielo nocturno está lleno de astros muy lejanos que no podemos observar a simple vista, sin embargo, formamos parte de una red cósmica llena de galaxias. La vía láctea es la galaxia a la que pertenecemos, nuestro sistema solar está rotando en uno de los brazos de esta galaxia. Nuestra vecina más cercana Andrómeda, una galaxia espiral situada a unos 2.5 millones de años luz de distancia, la cual se está aproximando a la vía láctea y ambas colisionarán en unos miles de millones de años, lástima que no vamos a estar presentes para observar este magnífico evento.

Los telescopios ópticos se han usado hasta la actualidad para ver astros, sin embargo, para ver más allá de lo que ven nuestros ojos, necesitamos de nuevas herramientas, las telecomunicaciones se desarrollan a pasos gigantescos. Los astrónomos empezaron a hacer observaciones astronómicas con antenas parabólicas a diferentes longitudes de onda. Con esta nueva tecnología se logra ver objetos mucho más lejanos y obtener información extra de lo que ocurre en nuestro universo.

En el último año, con el telescopio James Webb se obtuvieron imágenes infrarrojas del espacio profundo con una cantidad de datos inmensa. Una imagen de este tipo contiene información de muchísimas galaxias, por lo que a un ritmo constante nos llenaríamos de datos de galaxias rápidamente, por lo que sería de gran ayuda una forma más rapida de analizar todos estos datos. El estudio de estas imágenes es un nuevo reto, ya que una persona sería incapaz de obtener datos de cada galaxia. Con el uso de la inteligencia artificial se han logrado algoritmos para hacer esta tarea más sencilla. Por lo que una clasificación automática de galaxias es un buen inicio para el aprendizaje con estos datos.

Objetivos

Objetivo General

Realizar clasificadores clásicos automáticos de galaxias con una precisión mayor al 80 % con la ayuda de algoritmos de clasificación en deep learning para la manipulación de grandes datos de imágenes sobre galaxias.

Objetivos Específicos

- Construir clasificadores con diferente número de datos para observar su comportamiento mientras se aumenta el número de imágenes.
- Realizar clasificadores con datos equilibrados y con datos no equilibrados.
- Analizar el comportamiento en los diferentes algoritmos de clasificación ya que nuestra base de datos muestra una carga de información a la clase *Starforming*.
- Visualizar las características de las imágenes que hayan obtenido precisión mayor al 90 %.

Organización

Comenzamos con el primer capítulo, donde se explica la relación de este trabajo con las telecomunicaciones, de los instrumentos usados para la observación de galaxias y una breve explicación de los fenómenos físicos que se producen para la observación de astros, así como diferentes telescopios astronómicos y sus frecuencias de operación. En el segundo capítulo se abordan las características y propiedades de las galaxias, así como su clasificación por morfología y por longitud de onda que emiten. Se muestran ejemplos de imágenes reales de la base de datos usada, con la intención que el lector pueda distinguir entre imágenes de diferentes clases. En esta sección se concentran las figuras más impresionantes y llamativas de galaxias. Seguido del tercer capítulo, donde se presentan los temas de machine learning

para aprendizaje supervisado, la teoría sobre los clasificadores y el detalle matemático de los algoritmos de clasificación automática empleados. Una vez entendida toda la teoría de galaxias y aprendizaje supervisado, en el cuarto capítulo continuamos con los métodos empleados para realizar los clasificadores, así como las herramientas usadas y una explicación de la base de datos. Después de todo el proceso de los clasificadores y diferentes combinaciones, se muestran los resultados de la precisión en los diferentes clasificadores, además se muestran gráficas de evolución variando el número de datos dando resultados bastante interesantes. Por último, en el sexto capítulo se muestran las conclusiones y los trabajos a futuro.

Motivación

El desarrollo de este trabajo responde al interés de las comunicaciones con el espacio exterior a través de los telescopios, por el asombro del universo y los objetos tan interesantes que contiene. Más específicamente, por las formas y patrones que muestran los objetos del universo. De igual forma por el maravilloso mundo de la computación, la inteligencia artificial y el desarrollo de algoritmos con el uso de las matemáticas y los secretos que guardan los números.

Capítulo 1

Ondas electromagnéticas en el espacio

Las ondas electromagnéticas son un fenómeno físico a las cuales estamos muy acostumbrados: la luz que vemos todos los días, hornos de microondas, señales wi-fi, en general están presentes en todas las telecomunicaciones. Tan diverso es, que las podemos encontrar en los sitios más lejanos del universo. El espacio profundo es un lugar inmenso, donde las distancias son inimaginables, lo cual observarlo es un tanto difícil. En una noche estrellada podemos ver puntos de luz en el cielo, lo que son estrellas relativamente cercanas, si nos ponemos más interesantes, podemos pensar objetos mucho más lejanos, aquí se encuentran objetos que parecen sacados de películas de ciencia ficción, el problema de observar estos objetos tan lejanos es que la luz no llega con la intensidad necesaria para ser detectados en la tierra. Por suerte nuestro espectro electromagnético cuenta con diversas longitudes de onda, por lo cual estos objetos se pueden observar con diferentes longitudes de onda y obtener información a pesar de la distancia, todas estas ondas viajan a la velocidad de la luz, a unos trescientos mil kilómetros por segundo (3×10^8 m/s).

1.1. Espectro electromagnético

Es el conjunto continuo e infinito de ondas electromagnéticas, ordenadas en zonas en función de su longitud de onda y, por tanto, de la energía que transportan. De mayor a menor longitud de onda, se tienen: ondas de radio, microondas, infrarrojos, luz visible, luz ultravioleta, rayos X y rayos gamma. El análisis de la luz emitida o reflejada por los objetos en el espacio en relación a sus diferentes longitudes de onda o energías es de gran importancia puesto que constituye la mayor fuente de información sobre el universo.

La astronomía se encarga del estudio de estas ondas, hay diferentes estudios en función de la zona del espectro electromagnético que se analiza como son, por ejemplo, la astronomía de altas energías (rayos X y gamma), la astronomía ultravioleta, la astronomía óptica (luz visible), la radio astronomía y la astronomía infrarroja [10]. La Fig 1.1. Muestra el espectro electromagnético, que abarca desde las ondas de radio hasta los rayos gamma. Nota: Las figuras sin créditos son de autoría propia.

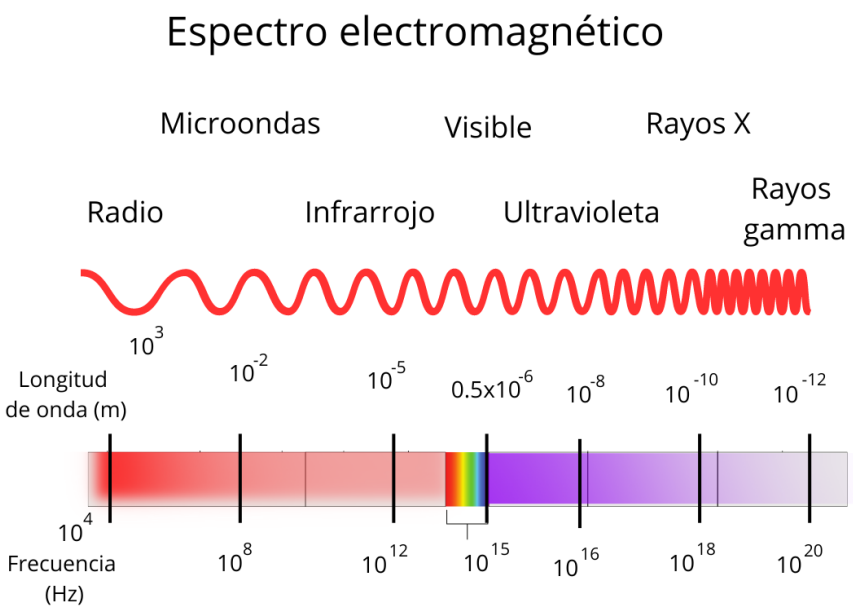


Figura 1.1: Espectro electromagnético.

1.1.1. Ondas infrarrojas

Nuestros ojos reciben longitudes de onda específicos de tal manera que nos indica el color de lo que nos rodea, sin embargo, una retina solo puede captar un rango bastante pequeño de todo el espectro electromagnético. Si avanzamos a menores frecuencias nos encontramos con la radiación infrarroja que se extiende desde el borde nominal rojo del espectro visible a 700 nanómetros (nm) hasta 1 milímetro (mm). Este rango de longitudes de onda corresponde a un rango de frecuencia de aproximadamente 430 THz hasta 300 GHz. Estas ondas son utilizadas para obtener información del universo que no podemos ver. La astronomía infrarroja utiliza telescopios equipados con sensores para penetrar en regiones polvorientas del espacio como las nubes en los centros de las galaxias, detectar objetos como nebulosas y fenómenos del universo temprano [10]. La Fig 1.2. Muestra "los pilares de la creación", nubes densas de gas y polvo donde se crean nuevas estrellas en la constelación de Serpens. tomada por el telescopio James Webb, que captura imágenes infrarrojas.

Figura 1.2: Los pilares de la creación. Créditos: NASA, ESA, CSA, STScI; Joseph DePasquale (STScI), Anton M. Koekemoer (STScI), Alyssa Pagan (STScI).

1.2. Efecto Doppler en ondas electromagnéticas

El efecto doppler es el cambio de frecuencia aparente de una onda producido por el movimiento relativo de la fuente respecto a su observador. Si el objeto se aleja, su luz se desplaza a longitudes de onda más largas, produciéndose un corrimiento hacia el rojo. Si el objeto se acerca, su luz presenta una longitud de onda más corta. Se ha utilizado para medir la velocidad a la que estrellas y galaxias están acercándose o alejándose de la Tierra. De esta forma se descubrió que la galaxia de Andrómeda se acerca a nuestra galaxia [10]. La Fig 1.3. Muestra el corrimiento al rojo y al azul producido por el efecto doppler.

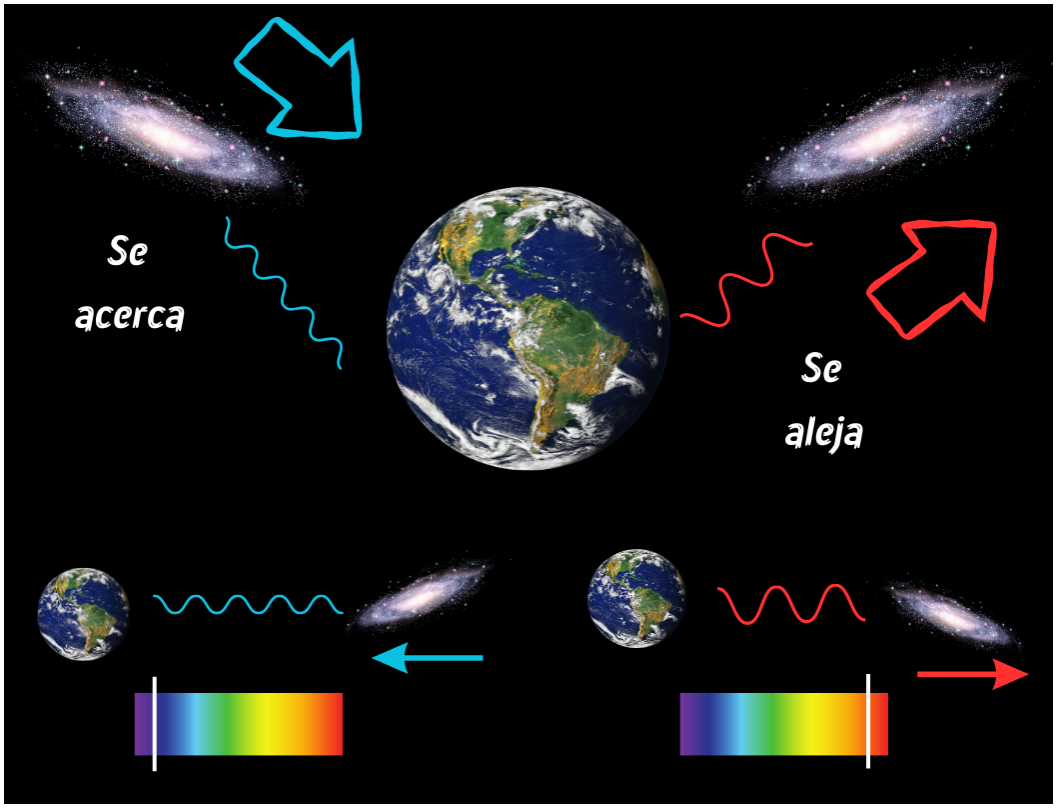


Figura 1.3: Efecto doppler en movimientos de galaxias.

Por este fenómeno se puede obtener información del universo profundo, entender los cúmulos de galaxias y la expansión del universo.

La ecuación para el corrimiento Doppler tanto con una fuente en movimiento como con un observador en moviemiento está dada por.

$$f' = f\left(\frac{v \pm v_0}{v \pm v_s}\right)$$

donde f' es la frecuencia recibida, f es la frecuencia original, v es la velocidad de la onda, v_0 es la velocidad del observador y v_s velocidad de la fuente. Los signos superiores de la ecuación se usan cuando el observador o la fuente se están acercando y los signos inferiores se utilizan cuando el observador o la fuente se están alejando [5].

Las ecuaciones anteriores son utilizadas para temas relacionados principalmente para ondas sonoras. Las ondas electromagnéticas también presentan un desplazamiento Doppler, excepto que la velocidad relativa entre la fuente y el receptor nunca puede ser mayor que la velocidad de la luz y la fórmula para calcular el corrimiento es ligeramente diferente. Para las ondas electromagnéticas tenemos:

$$f' = f\left(\frac{c+v}{c-v}\right)^{\frac{1}{2}}$$

Donde f' es la frecuencia recibida, f es la frecuencia original, v es la velocidad relativa entre el observador y la fuente y c es la velocidad de la luz. La ecuación anterior igual puede quedar de la siguiente forma:

$$f' = f\sqrt{\frac{c+v}{c-v}}$$

1.3. Telescopios

Los telescopios son herramientas que usan los astrónomos para observar objetos lejanos, los telescopios ópticos funcionan enfocando la luz utilizando espejos curvos. El telescopio Hubble es de este tipo, llamado así en honor a Edwin Hubble. Muchas de las imágenes observadas en este documento son obtenidas de este telescopio. Las observaciones ópticas tienen sus limitantes, siempre están sujetos a la cantidad de luz que emite el objeto y otras materiales que pueden disminuir los fotones obtenidos, el polvo y nubes de gas pueden distorsionar la luz que llega al planeta tierra. Con el avance de la tecnología se crearon nuevos instrumentos de observación que trabajan en diferentes frecuencias que no se pueden observar con nuestros ojos. Con estas longitudes de onda se obtiene mucho más información, ya que atraviesan el polvo. Estos telescopios tienen forma de parábola para obtener la mayor información de datos y concentrarla en un punto, por lo que se usan antenas parabólicas.

1.3.1. Antenas parabólicas

Las antenas son dispositivos con el objetivo de recibir y/o emitir ondas electromagnéticas por el espacio libre, se encargan de convertir energía eléctrica en ondas electromagnéticas, también son capaces de realizar el proceso inverso. Las antenas parabólicas son usadas para la recepción de canales de radio y televisión, así son llamadas ya que tienen forma de parábola y tienen la propiedad que las señales se concentran en un foco. Este tipo de antenas se han usado desde hace muchos años por los astrónomos para captar las ondas electromagnéticas del espacio exterior. De este modo se han obtenido imágenes del universo profundo, formando una nueva rama de la astronomía llamada radioastronomía.

1.3.2. Antenas en Tulancingo

En el estado de Hidalgo se realizan proyectos ambiciosos de astronomía. En la ciudad de Tulancingo se inauguró la Estación Terrena para comunicaciones vía satélite en 1968. Las antenas parabólicas de Tulancingo fueron diseñadas para recibir y transmitir señales entre la tierra y los satélites artificiales, por ellas se transmitieron los juegos olímpicos XIX. En la actualidad se proyecta convertir una de estas antenas de 32 metros en un radiotelescopio. Tiene un lugar privilegiado al ubicarse entre una cadena montañosa que lo protege de interferencia de microondas. Se busca observar objetos mucho más lejanos como galaxias y estrellas. Con este proyecto se pretende trabajar en frecuencias más bajas y colaborar con otros países, así como se ha hecho con el Gran Telescopio Milimétrico. La Fig 1.4. Muestra la antena TIL-1, antena parabólica que se planea usar para observaciones de galaxias y estrellas lejanas.

Figura 1.4: Antena TUL-1 en Tulancingo. Créditos: SCT, Secretaría de Comunicaciones y Transporte.

1.3.3. Gran Telescopio milimétrico

El gran telescopio milimétrico está situado en Puebla México, dentro del Parque nacional del pico de Orizaba. Este telescopio que se observa en la Fig. 1.5. Es el más grande y potente de su clase en el mundo, opera con longitudes de onda de hasta un milímetro, a cargo del Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE). Estudia los fenómenos relacionados con la formación de galaxias, estrellas y planetas. Cuenta con una gran antena alt-azimutal de altísima precisión de 50m de diámetro, que opera entre los 75 y 300 GHz. Una de las principales áreas de investigación del GTM es el estudio del universo primitivo. Principalmente el estudio de galaxias de núcleo galáctico activo (AGN) donde su radiación es mayormente infrarroja y donde se desplaza su emisión al rango milimétrico y submilimétrico. En 2019 formó parte de la primera imagen de un agujero negro situado en el centro de nuestra galaxia.

Figura 1.5: Gran telescopio milimétrico.

1.3.4. Bandas de telecomunicaciones

En los Estados Unidos Mexicanos el Instituto Federal de Telecomunicaciones es el ente encargado de regular y supervisar las redes y la prestación de servicios de telecomunicaciones y radiodifusión. El espectro radioeléctrico es un bien de dominio público, por lo que están habilitadas ciertas porciones del espectro para comunicaciones inalámbricas. Estas porciones se usan para diferentes propósitos, el uso de radares, comunicaciones AM, señal de televisión y celular, transmisión satelital y las bandas donde pueden operar los radiotelescopios. El GTM situado en el estado de Puebla, opera en la banda EHF y no se permite la operación de ningún sistema de comunicación dentro de un área de 100 km alrededor de él, ya que puede causar interferencia en las señales. Sin embargo, los radiotelescopios pueden funcionar en diferentes bandas. En cada banda se puede recibir diferentes tipos de señales. Estas bandas están asignadas según su frecuencia y la longitud de ondas que opera cada banda, cada banda tiene propiedades y se usan para diferentes actividades de comunicaciones. La Fig 1.6. Muestra las bandas que son usadas para comunicaciones.

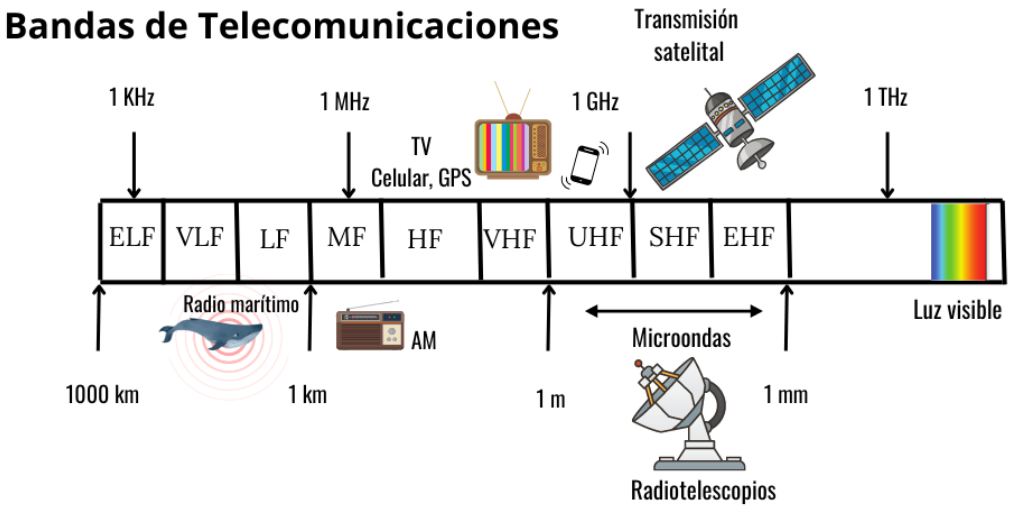


Figura 1.6: Bandas de telecomunicaciones.

1.3.5. Sloan Digital Sky Survey

El Sloan Digital Sky Survey (SDSS) es uno de los estudios más ambiciosos en la historia de la astronomía. Durante ocho años se obtuvieron imágenes profundas y multicolores que cubrían más de una cuarta parte del cielo con imágenes que contenían más de 930.000 galaxias, 120,000 cuásares y 225,000 estrellas.

El SDSS utilizó un telescopio dedicado de 2,5 metros en el Observatorio Apache Point, Nuevo México, equipado con dos potentes instrumentos especiales. La cámara de 120 megapíxeles capturó 1,5 grados cuadrados de cielo a la vez, unas ocho veces el área de la luna llena. Un par de espectrógrafos alimentados por fibras ópticas midieron espectros de cientos de miles de galaxias. En Fig 1.7. Se observa el telescopio utilizado para tomar las imágenes procesadas en este trabajo.

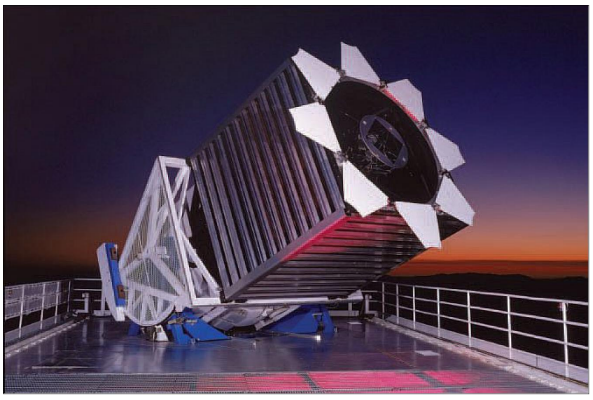


Figura 1.7: Telescopio SDSS de noche. créditos SDSS/ Patrick Gaulme.

Los datos SDSS han respaldado el trabajo fundamental en una extraordinaria variedad de disciplinas astronómicas, incluidas las propiedades de las galaxias, la evolución de los cuásares, la estructura y las poblaciones estelares de la Vía Láctea.

El telescopio del SDSS se encuentra en el Observatorio de Apache Point (APO) en Sunspot, Nuevo México. El observatorio está rodeado por el Bosque Nacional Lincoln en las montañas Sacramento y se sitúa en una montaña a 2800 metros sobre el nivel del mar, donde la atmósfera contiene una cantidad escasa de vapor de agua y muy pocos contaminantes [11].

Capítulo 2

Galaxias

Las galaxias son un sistema de estrellas, gas y polvo unidos por la fuerza de gravedad. Las galaxias son los componentes fundamentales del universo. Algunas de ellas tienen una estructura sencilla, contienen estrellas normales y no muestran características particulares. También hay galaxias que están casi completamente formadas por gas neutro [5]. Por otro lado, otras son sistemas complejos construidos a partir de muchos componentes complejos de estrellas, polvo, nubes moleculares, campos magnéticos y rayos cósmicos [12]. Las galaxias pueden formar pequeños grupos o grandes cúmulos en el espacio. La Fig 2.1. Muestra la galaxia espiral NGC 5037, que se encuentra a unos 150 millones de años luz en la constelación de Virgo.

Figura 2.1: Galaxia espiral NGC 5037. Créditos (ESA/Hubble & NASA, D. Rosario; Acknowledgment: L. Shatz).

2.1. Características de las galaxias

En el centro de muchas galaxias hay un núcleo compacto que a veces puede ser tan brillante que abruma toda la radiación normal de la galaxia. La luminosidad de las galaxias normales más brillantes pueden corresponder a 10^{12} luminosidades solares, pero la mayoría de ellas son mucho más débiles. Una luminosidad solar es la radiación o energía emitida por nuestro sol, que asciende a casi cuatrocientos cuatrillones de vatios [8]. La densidad de la materia puede ser muy diferente en distintas galaxias y en diferentes partes de la misma galaxia. La evolución de una galaxia será el resultado de procesos que ocurren en escalas de tiempo y energía muy diferentes [5]. La Fig. 2.2. Muestra galaxias espirales obtenidas de la página oficial de la NASA.

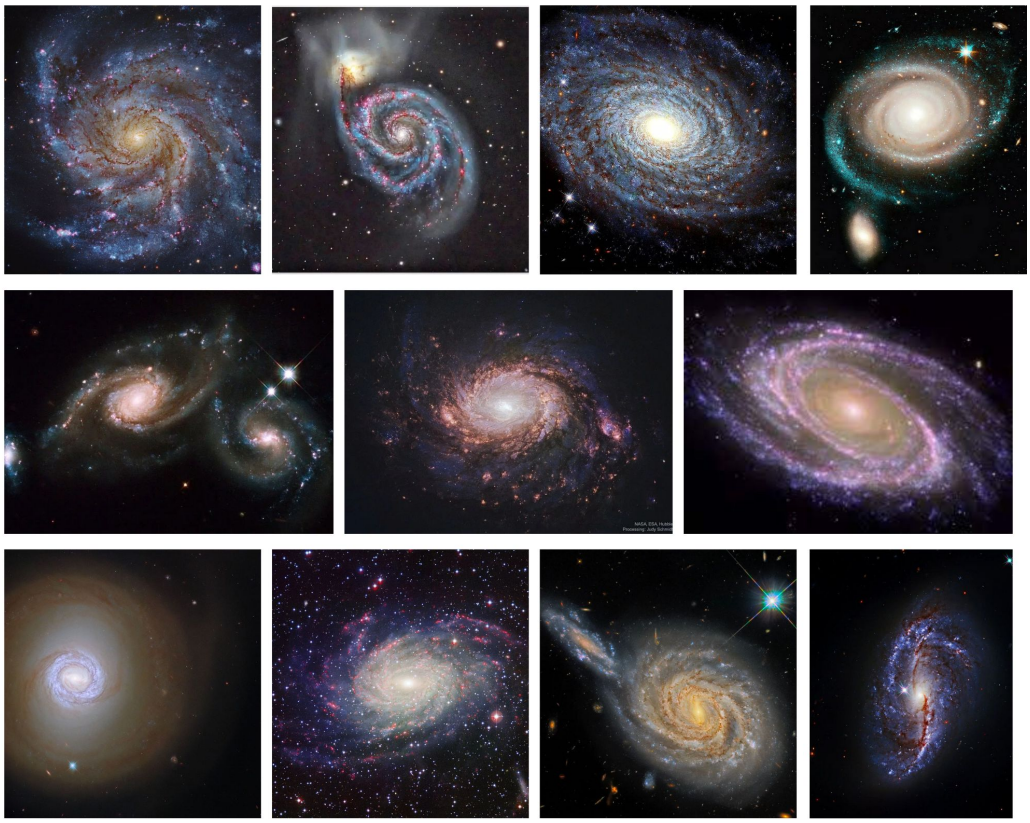


Figura 2.2: Galaxias espirales. `https://www.nasa.gov/multimedia/imagegallery`.

2.2. Clasificación de galaxias

Un primer paso útil para comprender las galaxias es una clasificación basada en sus diversas formas. Aunque una clasificación morfológica siempre debe ser subjetiva. Sin embargo, hay que considerar que las imágenes obtenidas estarán limitadas a aquellas galaxias que sean lo suficientemente grandes y brillantes para ser fácilmente visibles en el cielo. Para que una clasificación sea útil, debe corresponder aproximadamente a las propiedades físicas importantes de las galaxias. La mayoría de las clasificaciones coinciden en sus principales características con la propuesta por Edwin Hubble en 1926. La secuencia de Hubble se muestra en la Fig 2.3.

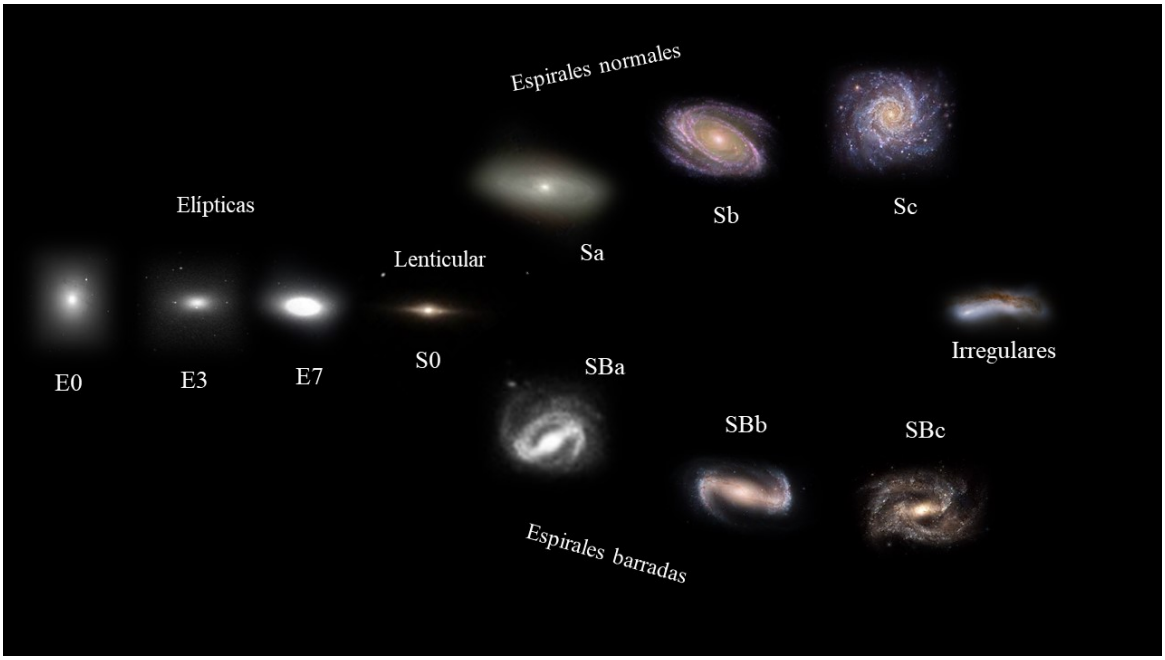


Figura 2.3: Clasificación de Hubble.

Los distintos tipos de galaxias están ordenados en una secuencia de tipos tempranos a tardíos. Se muestran tres principales tipos de galaxias: Elípticas, Lenticulares y Espirales. Las espirales se dividen en dos secuencias, espirales normales y barradas. Además, Hubble incluyó una clase de galaxias irregulares. En la secuencia de Hubble, las galaxias lenticulares o S0 se sitúan entre las elípticas y espirales.

2.3. Clasificación por morfología

2.3.1. Elípticas

Las galaxias elípticas se caracterizan por ser suaves distribuciones elípticas de brillo superficial, en el que la densidad disminuye de forma regular a medida que se avanza hacia el exterior. Se conocen como galaxias caníbales, ya que son las más grandes y por lo tanto, las más masivas. Son de edad estelar viejas y poco gas. Este tipo de galaxias no están girando a grandes velocidades, por lo tanto pueden tener esa forma elíptica y tienen colores fotométricos rojos, características de una población estelar antigua. Están divididas en subtipos $E_0, E_1, ...E_7$, su tipo se define como E_n, donde $n = 10(1 - b/a)$ con a y b como las longitudes de los semiejes mayor y menor. Por lo tanto, una galaxia E_0 tiene un aspecto circular en el cielo [8]. La Fig 2.4. Muestra una galaxia elíptica en la constelación Ursa mayor a unos 80 millones de años luz de distancia, descubierta en 1793 por William Herschel.

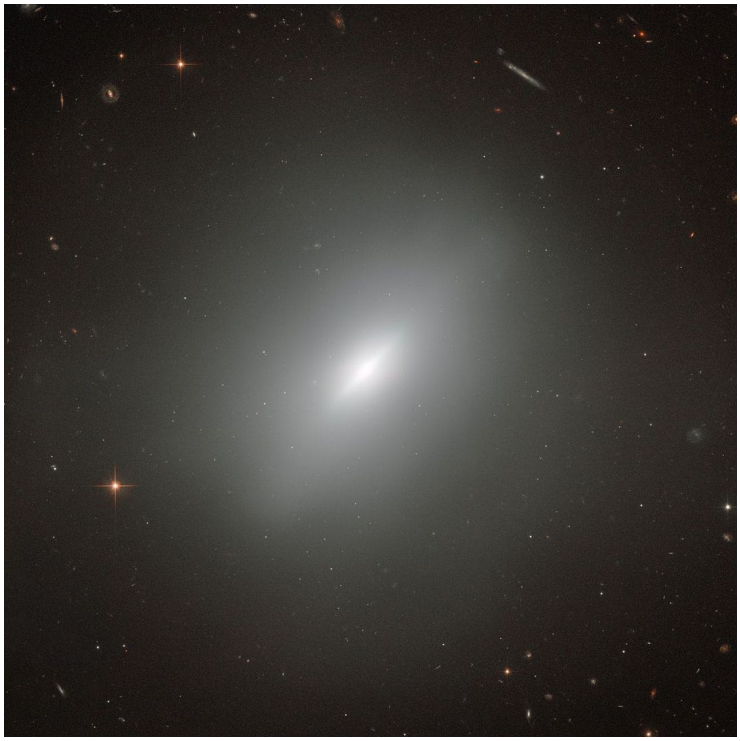


Figura 2.4: Galaxia elíptica NGC 3610. Créditos: ESA/Hubble & NASA, Acknowledgement: Judy Schmidt.

2.3.2. Espirales

El rasgo característico de las galaxias espirales es un patrón espiral más o menos bien definido en el disco. Normalmente tienen dos brazos, aunque existen casos con mayores números de brazos. Una galaxia espiral consiste en un bulbo central que es estructuralmente similar a una galaxia E, y de un disco estelar, como en una galaxia SO. Además, hay un disco delgado de gas y otra materia interestelar [12]. Son galaxias pequeñas con poca densidad y ricas en helio e hidrógeno. Estas galaxias son más azules que las galaxias elípticas de la misma luminosidad, principalmente por estrellas gigantes azules por formación estelar reciente. Son las más abundantes del universo. Hay dos secuencias espirales: Las normales Sa-Sb-Sc y barradas SBa-SBb-SBc [8].

Espirales normales

En las espirales normales el patrón puede terminar en un anillo interior o continuar hasta el centro. La Fig. 2.5. Muestra una galaxia espiral, en el extremo de un brazo se observa una interacción con una galaxia espiral barrada más pequeña.

Figura 2.5: Galaxia espiral NGC 1378. Créditos: NASA, ESA, The Hubble Heritage Team (STScI/AURA), and A. Riess (JHU/STScI).

Espirales barradas

Estas galaxias tienen en la parte central lo que se conoce como barra, tienen un núcleo que no es esférico, más bien es elongado. Los brazos emergen a partir de esta barra. Nuestra galaxia la vía láctea es una galaxia espiral barrada típica. En la Fig. 2.6. Se muestra la galaxia espiral barrada NGC 1300 que se encuentra a unos 70 millones de años luz a orillas de la constelación de Eridanus.

Figura 2.6: Galaxia espiral barrada NGC 1300. Créditos: STScI-2005-01. NASA, ESO.

2.3.3. Lenticulares

En la secuencia de Hubble, las galaxias lenticulares o SO se sitúan entre las elípticas y espirales. Al igual que las elípticas, contienen poca materia interestelar y no muestran signos de una estructura espiral. Sin embargo, además del componente estelar elíptico habitual, también contienen un disco plano formado por estrellas [10]. La Fig. 2.7. Muestra la galaxia NGC 5010 de tipo S0-a, situado a unos 136 millones de años luz de nuestro sistema solar en la constelación de la virgen. Se considera una galaxia con luminosidad infrarroja, está pasando de ser una galaxia espiral a una elíptica, con sus brazos espirales quemados y convertidos en brazos polvorientos. Desde la perspectiva de la tierra, la galaxia se muestra casi de canto.

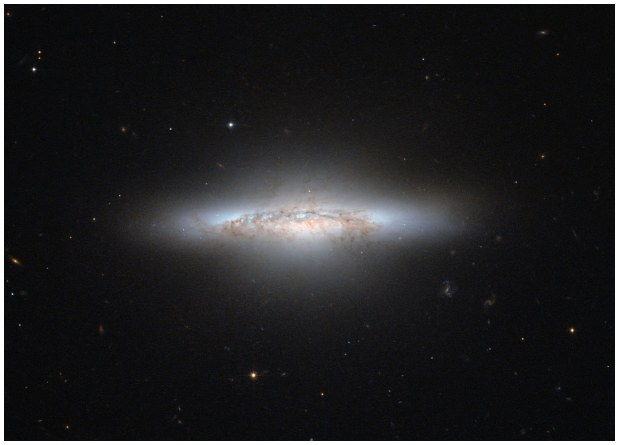


Figura 2.7: Galaxia lenticular NGC 5010 .Créditos: ESA/Hubble & NASA.

2.3.4. Irregulares

Estos objetos no tienen un bulbo dominante ni un disco rotacionalmente simétrico y carecen de una simetría evidente. Son ricas en gas y contienen muchas estrellas jóvenes [10]. La Fig. 2.8. Muestra la galaxia irregular NGC 520, situada a 100 millones de años luz, en la constelación de Picis. Descubierta en 1784 por William Herschel, se considera que son dos galaxias, posiblemente espirales interactuando entre sí y posiblemente colisionando vistas de lado.

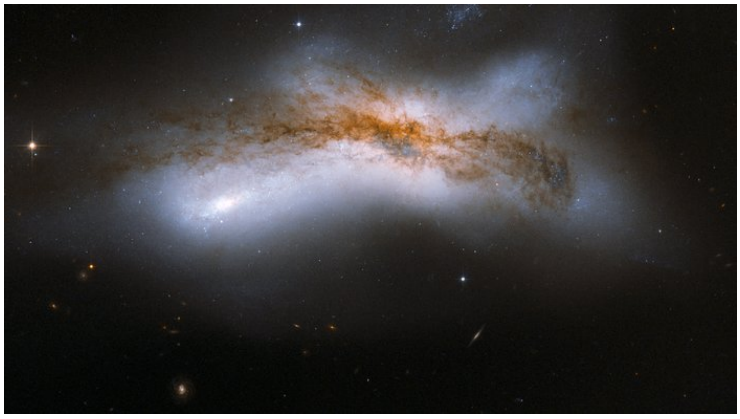


Figura 2.8: Galaxia irregular NGC 520. Créditos: NASA, ESA, the Hubble Heritage Team (STScI/AURA)-ESA/Hubble Collaboration and B. Whitmore (STScI).

Esta clasificación de galaxias muestra patrones que serán procesados por los algoritmos, sin embargo, en algunas clases estos patrones no son tan claros, por lo que clasificar estas imágenes será un reto.

2.4. Clasificación por longitud de onda

Las galaxias emiten radiación electromagnética, relacionada con una longitud de onda específica. Al analizar las diferentes longitudes de onda, se obtiene información de la galaxia que no podría ser posible si únicamente se analizara por lo que podemos observar con luz visible. El espectro que emiten las galaxias es la suma del espectro que emiten todos los objetos que conforman a las galaxias. Estos espectros son analizados desde la tierra y con esta información podemos clasificar las galaxias [1]. La Fig. 2.9. Muestra nuestra vecina más próxima, la galaxia de Andrómeda vista desde diferentes longitudes de onda.

Figura 2.9: Galaxia Andrómeda vista desde diferentes longitudes de onda. Créditos: infrared: ESA/Herschel/PACS/SPIRE/J. Fritz, U. Gent; X-ray: ESA/XMM-Newton/EPIC/W.

Los astrónomos obtienen mucha información a partir de analizar los espectros de estrellas, galaxias y cuásares. El objetivo de la espectroscopía es obtener las distribuciones espectrales respecto a la longitud de onda. La espectroscopía permite distinguir componentes de las galaxias (estrellas, gas caliente, polvo, etc).

La información que se obtiene con la espectroscopía es mucho mayor que con la fotometría. Permite clasificar directamente las estrellas. Se obtienen datos de la medida de sus líneas, dando temperaturas y abundancias de elementos en la atmósfera, rotación, velocidad de desplazamiento respecto al observador etc [13].

Para la clasifiación de galaxias por espectroscopía se usa el fenómeno del desplazamiento al rojo causado por el efecto doppler. El desplazamiento al rojo es una medida del cambio de la longitud de onda de una línea conocida del espectro de un objeto astronómico.

Cada objeto espectroscópico tiene una clasificación (clase) y una determinación de desplazamiento al rojo. Para las galaxias, se puede determinar una dispersión de velocidad (hasta unos 70 km/s). Los corrimientos al rojo se determinan ajustando modelos a cada espectro asumiendo una amplia gama de posibles corrimientos al rojo. Los datos del SDSS que describen las longitudes de onda de la línea espectral utilizan longitudes de onda de vacío [11]. Sin embargo, las longitudes de onda de las transiciones atómicas generalmente se consideran a temperatura y presión estándar. Por lo tanto, reconocer líneas espectrales asociadas con transiciones atómicas puede requerir convertir los datos SDSS a los valores equivalentes en STP. La tabla 2.1. Es una conversión de longitudes de onda de aire a vacío dada su línea espectral.

Línea	Aire	Vacio
H-Beta	4861.363	4862.721
[OIII]	4958.911	4960.295
[OIII]	5006.843	5008.239
[NII]	6548.05	549.86
H-alfa	6562.801	6564.614
[NII]	6583.45	6585.27
[S II]	6716.44	6718.29
[S II]	6730.82	6732.68

Tabla 2.1: Longitud de onda en el aire y vacío de algunas transiciones comunes.

2.4.1. Starforming

La formación de estrellas en las galaxias espirales genera grandes cantidades de luz ultravioleta que es absorbida por el polvo y re-irradiada en longitudes de onda infrarrojas.

Se asigna esta clase en función de si la galaxia tiene líneas de emisión detectables que son consistentes con la formación estelar de acuerdo con los criterios:

$$log_{10}(OIII/H\alpha) < 0.7 - 1.2(log_{10}(NII/H\alpha) + 0.4)$$

La Fig. 2.10. Muestra el objeto Messier 51, conocida como galaxia de remolino, de tipo starforming. Está situada a 31 millones de años luz de la tierra en la constelación de Canes Venatici. Sus brazos son fábricas de formación estelar, que comprimen gas hidrógeno y crean cúmulos de nuevas estrellas. El color rojo representa la luz infrarroja así como el hidrógeno dentro de las regiones de formación de estrellas gigantes. El color azul se puede atribuir a estrellas jóvenes y calientes, mientras que el color amarillo es de estrellas más viejas [12]. En el extremo de uno de sus brazos hay una interaccion con la galaxia NGC 5195.

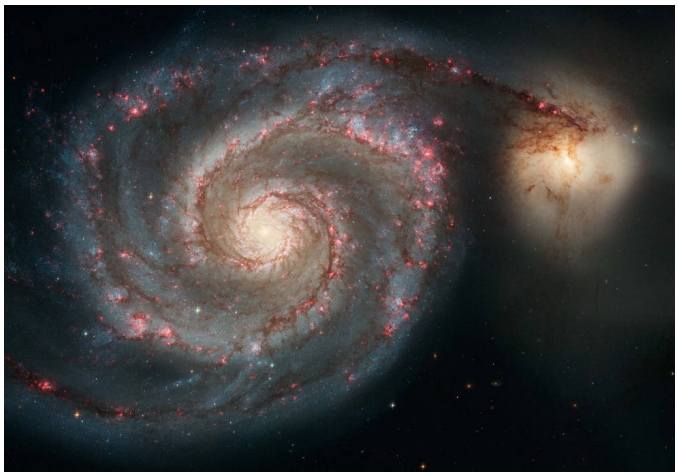


Figura 2.10: Galaxia de remolino, también conocida como messier 51. Créditos: S. Beckwith (STScI)Hubble Heritage Team, (STScI/AURA), ESA, NASA.

En nuestra base de datos la clase Starforming tiene imágenes con un patrón de parecer galaxias de disco con espirales y un centro brillante. A continuación se muestran ejemplos de imágenes de la clase starforming. Como se puede observar tienen una forma en espiral y color azul en sus brazos espirales, así como un centro brillante en su centro.

La Fig. 2.11 muestra imágenes reales de la clase starforming de nuestra base de datos empleada.

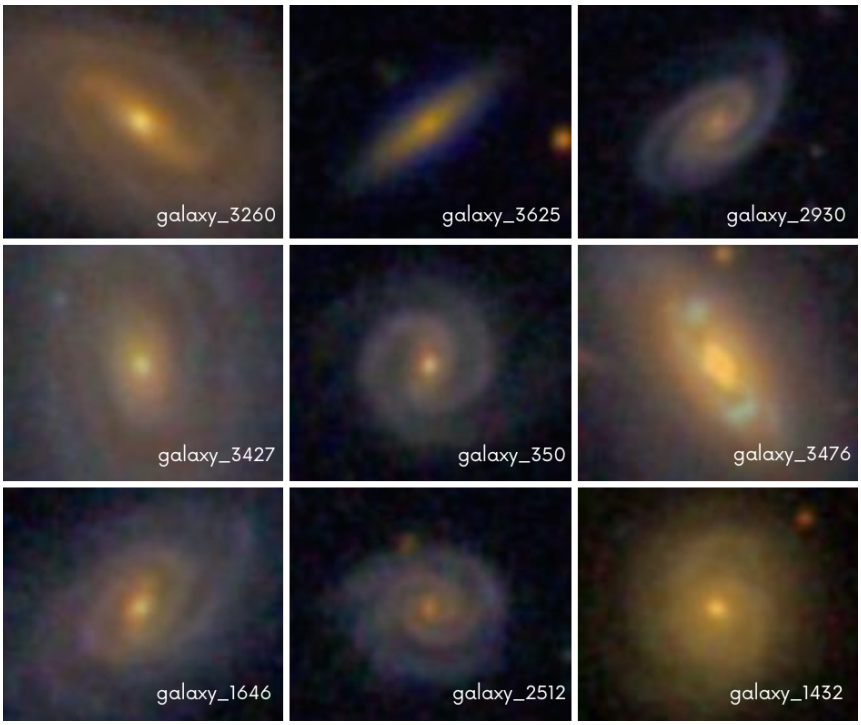


Figura 2.11: Imágenes de galaxias Starforming.

2.4.2. Broadline

Las líneas de emisión anchas se observan en el espectro de muchos AGN's. Suelen tener grandes velocidades en los núcleos de las galaxias [9]. La Fig. 2.12. Muestra un luminoso cuásar, relacionado con los broadline.

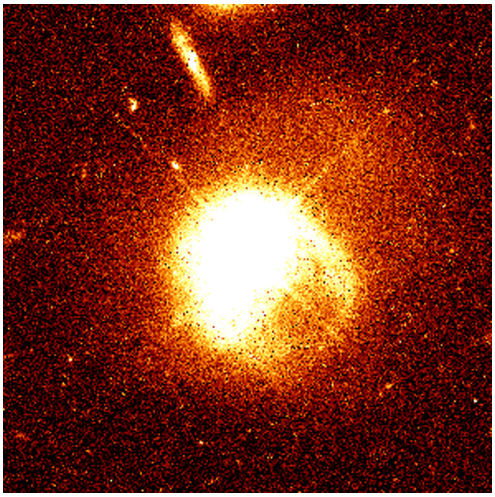


Figura 2.12: Cuásar PKS- 2349. Créditos: John Bahcall (Institute for Advanced Study, Princeton) Mike Disney (University of Wales) and NASA/ESA.

Si alguna galaxia tiene líneas detectadas en el nivel 10− sigma, con sigmas $> 200km/seg$ en el nivel $5 - sigma$, se agrega la indicación "BROADLINE" a su subclase.

La Fig. 2.13. Muestra ejemplos de imágenes de la clase Broadline, se puede observar galaxias con forma de disco, con espirales menos claras y centros brillantes, muestran un color amarillo en la mayoría de las galaxias.

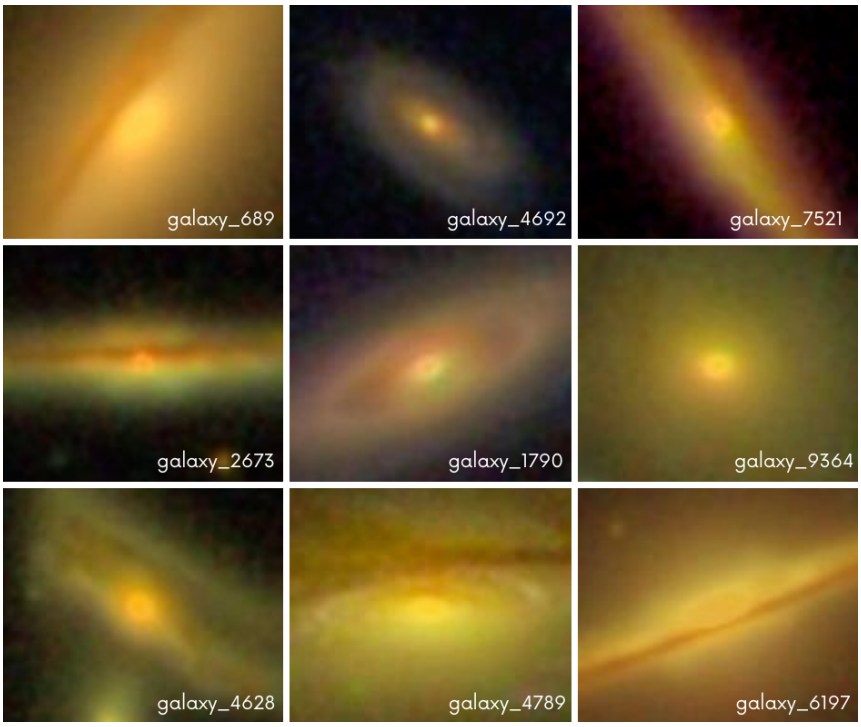


Figura 2.13: Imágenes de galaxias Broadline.

2.4.3. AGN

En el núcleo de algunas galaxias se produce una violenta actividad, que se denominan galaxias de núcleo galáctico activo (AGN). Las luminosidades de los núcleos galácticos activos pueden ser extremadamente grandes, a veces mucho mayor que del resto de la galaxia [9]. La Fig. 2.14. Muestra una imágen de una galaxia AGN, Centaurus A, situada a unos 13.05 millones de años luz.

Figura 2.14: Galaxia de núcleo activo Centaurus A. Créditos: X-ray: NASA/CXC/SAO; Optical: Rolf Olsen; Infrared: NASA/JPL-Caltech.

Parece poco probable que una galaxia pueda mantener una producción de energía tan grande durante mucho tiempo. Por esta razón se piensa que las galaxias activas representan una etapa pasajera en la evolución de las galaxias normales [12].

Se clasifican en función de si la galaxia tiene líneas de emisión detectables que son consistentes con ser un Seyfert o LINER:

$$log_{10}(OIII/H\alpha) > 0.7 - 1.2(log_{10}(NII/H\alpha) + 0.4)$$

En la Fig. 2.15. Se observan imágenes de clase AGN de la base de datos.

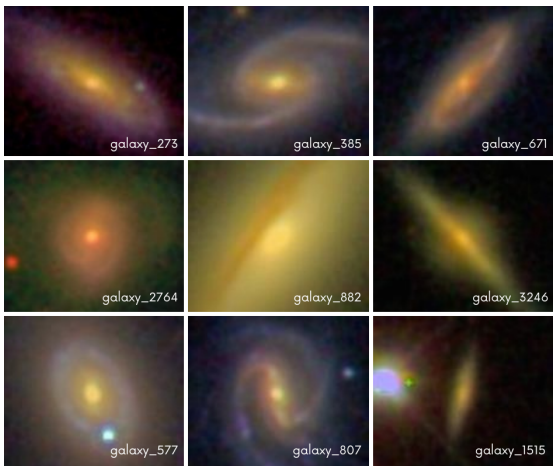


Figura 2.15: Ejemplos imágenes de galaxias AGN.

2.4.4. Starburst

Este tipo de galaxias son extremadamente brillantes en las longitudes de onda infrarrojas, mostrando hasta un 98 % de su luminosidad en el infrarrojo [1]. Son más brillantes en el infrarrojo que en el espectro óptico porque la luz visible es absorbida por las grandes cantidades de gas y polvo, el polvo vuelve a emitir energía térmica en el espectro infrarrojo.

La mayor parte de la formación estelar está confinada en el centro de la galaxia, se han detectado vastas nubes de hidrógeno que sirve de combustible. La Fig. 2.16. Muestra la galaxia NGC 3034 o también conocida como Messier 82, situada a aproximadamente 12 millones de años luz de nuestro sistema solar.

Figura 2.16: Galaxia starburst NGC 3034, créditos: National Astronomical Observatory of Japan (NAOJ).

La Fig. 2.17. Muestra imágenes de la clase starburst de la base de datos.

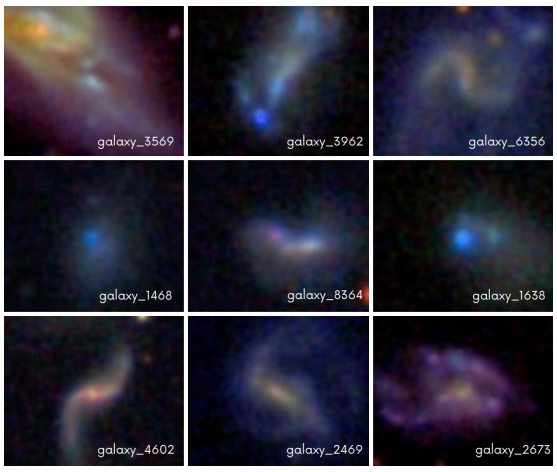


Figura 2.17: Ejemplos de imágenes de galaxias Starburst.

2.4.5. Ejemplos de base de datos

Después de esta explicación de las características de la clasificación por longitud de onda, la Fig. 2.18. Muestra ejemplos de todas las clases de nuestra base de datos. Se puede observar un color azul y sin forma definida para las galaxias starburst, un color amarillo o azul y forma de punto para las galaxias AGN, las clases starforming y broadline, son en su mayoria galaxias en forma espiral, el color azul predomina para las galaxias starforming y el color amarillo para las broadline.

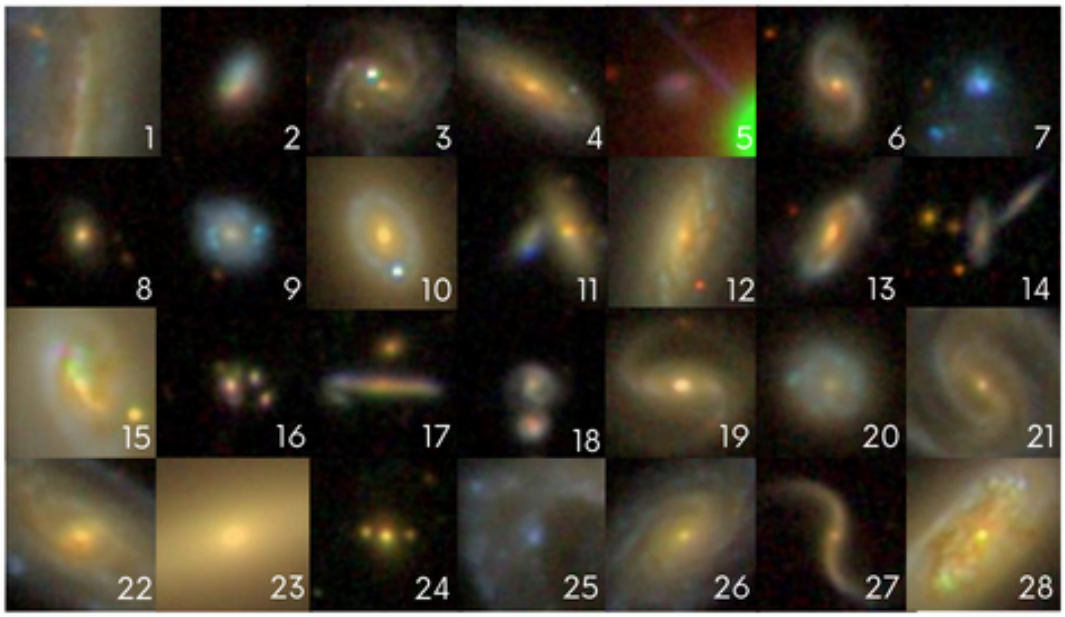


Figura 2.18: Imágenes de galaxias.

1. Starforming	8. Starforming	15. Starburst	22. AGN
2. Starforming	9. Starforming	16. Starburst	23. Starforming
3. Starforming	10. A GN	17. Starforming	24. AGN
4. Starforming	11. Starforming	18. Starforming	25. Starburst
5. Starforming	12. AGN	19. Starforming	26. Starburst
6. Starforming	13. Starforming	20. Starforming	27. Starforming
7. Starburst	14. Starforming	21. Starforming	28. Starburst

Capítulo 3

Aprendizaje Supervisado

3.1. Clasificadores

Un clasificador es un sistema cuya función es incluir en una clase dada, ya conocida, un elemento u observación nueva. Clasificar un objeto desconocido consiste en asignarlo a la clase en la cual las características usadas durante el entrenamiento tienen más correspondencia con las características del objeto [7]. Los procesos de modelos de clasificación se dividen en tres pasos importantes:

Entrenamiento del modelo. Un porcentaje de nuestra base de datos es asignado para el entrenamiento del modelo, para este trabajo se usa un 80 % del dataset. Este porcentaje de imágenes se eligen de manera aleatoria de la base de datos, por lo que hay corridas donde la precisión es mucho mayor ya que hay imágenes seleccionadas donde el modelo puede encontrar patrones de manera más eficiente que en otras corridas con diferentes imágenes, para solucionar esto se usa la validación cruzada o también conocido como *cross validation*, que veremos más adelante. Con estos datos el modelo aprende a reconocer patrones que tiene cada clase.

Prueba del modelo. Una vez entrenado el modelo, se pone a prueba. Para este paso se selecciona el resto de imágenes del dataset, para este trabajo se usa el 20 % restante. Con el modelo ya entrenado, se clasifican las imágenes, se muestra una precisión otorgada

por una prueba final donde se comparan los resultadas del modelo y las clases correctas a las que pertenece cada imagen, por lo que es importante que la base de datos ya esté clasificada, por esta razón del nombre de *Aprendizaje supervisado* ya que el modelo busca patrones en las clases y las compara con la información ya dada.

Explotación del modelo. Una vez obtenidos los resultados de la precisión se puede determinar si nuestro modelo se puede usar para explotar. Para que un clasificador se considere bueno, debe tener una precisión mayor de 90 %.

Una vez dicho esto, clasificar galaxias, no es más que asignar a cada imagen de la base de datos una clase ya determinada la cual debe tener ciertos patrones detectables en la imagen para pertenecer a cada clase. De tal modo que nuestro clasificador queda como se muestra en la Fig. 3.1.

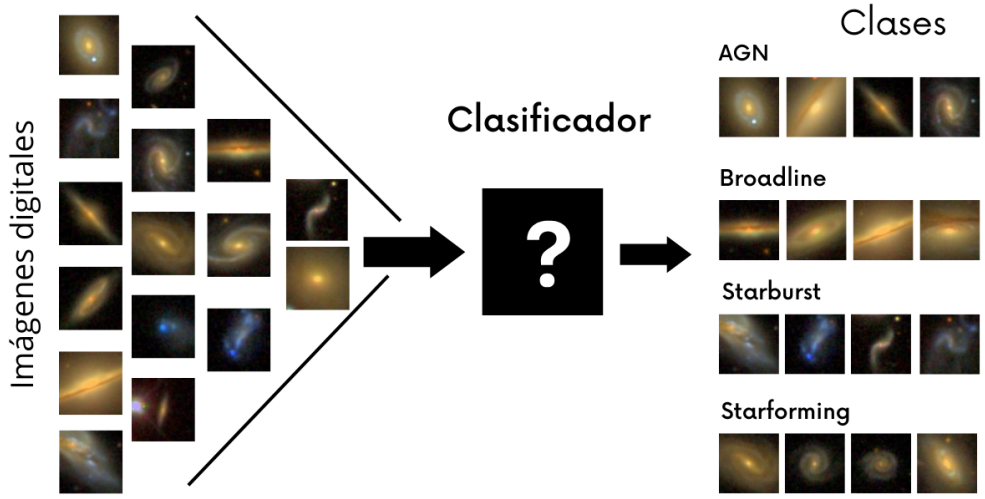


Figura 3.1: Diagrama de clasificador de galaxias.

En la parte central de la figura podemos observar un cuadro negro con un signo de interrogación; dentro de él ocurren todos los procesos matemáticos e informáticos que clasifican imágenes, podriamos poner a una persona a clasificar manualmente sesenta mil imágenes y obtener una clasificación intuitiva por las observaciones que realizó la perso-

na. Sin embargo, este método no resulta conveniente para la clasificación y la persona encargada, más bien, aquí adentro ocurren procesos un tanto complejos que se encargan de aprender de los patrones observados en las imágenes. Estos procesos pueden ir desde medir distancias euclidianas hasta la creación de complejas redes neuronales.

La clasificación empleada en este trabajo es por longitud de onda que emiten diferentes galaxias. Dado que es mucha la información en la base de datos, necesitamos de métodos que pueden realizar este tipo de procesamiento automáticamente. A continuación se muestran a detalle los cuatro algoritmos matemáticos empleados para clasificar las diferentes imágenes digitales.

3.2. Clasificadores empleados

Para este trabajo se hizo uso de cuatro algoritmos de clasificación automática clásicos, los cuales son descritos con detalle a continuación.

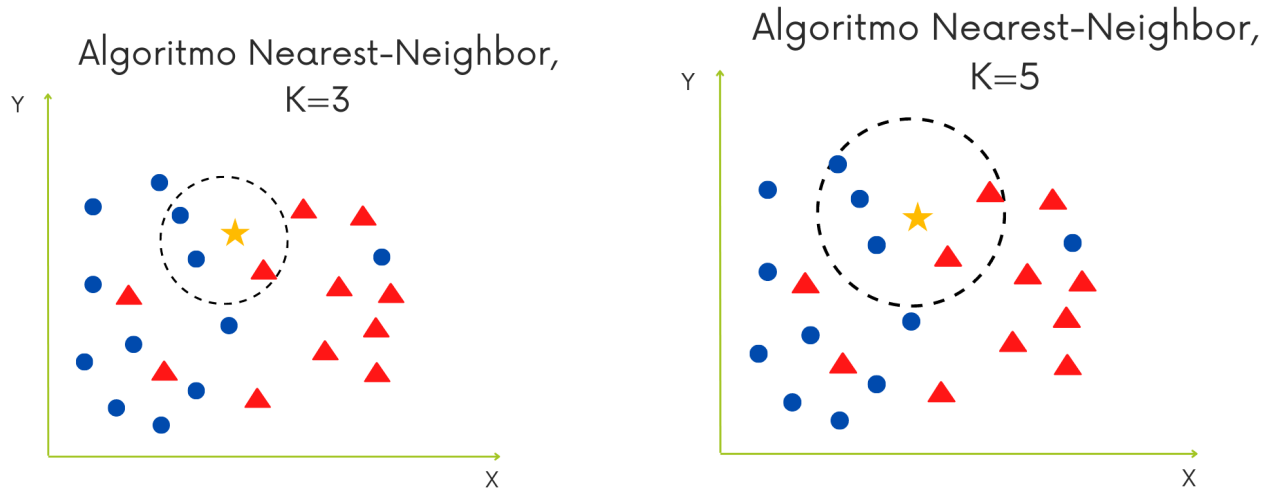
3.2.1. K- vecinos más cercanos (K-NN)

El clasificador K-nearest neighbors o K-vecinos más cercanos, es uno de los métodos de clasificación más sencillos y naturales, obtiene las etiquetas de clase de los nuevos objetos de entrada a partir de los ejemplos de entrenamiento más similares. La construcción de este clasificador es un caso especial de aprendizaje basado en instancias, este algoritmo no aprende del modelo, memoriza las instancias de formación que posteriormente se utilizan como conocimiento para la fase de predicción. La predicción se calcula por la mayoría de votos de los vecinos más cercanos [14].

El número de vecinos k es un hiperparámetro que se debe elegir en el momento de la construcción del modelo. Se puede pensar en k como una variable de control para el modelo de predicción. No existe un número óptimo de vecinos que se adapte a todo tipo de conjuntos de datos, cada conjunto de datos tiene sus propios requisitos. Una pequeña cantidad de vecinos tendrán un bajo sesgo, pero una alta varianza, y un gran número de

vecinos tendrán una varianza más baja, pero un sesgo más alto [3].

En la Fig. 3.2. Se observa la variación gráfica con diferentes números de vecinos, se observa como se va formando un circulo con radio cada vez mayor según se aumenta el número de vecinos.



(a) 3 vecinos cercanos

(b) 5 vecinos cercanos

Figura 3.2: Representación gráfica aumentando K.

En general, cualquier algoritmo de vecinos más cercanos tiene cuatro ingredientes:

Distancia métrica

Mide la distancia entre puntos. Determina cuáles de los ejemplos de entrenamiento están más cerca de un punto de datos de consulta, y selecciona el ejemplo de entrenamiento para calcular una predicción, normalmente para medir se hace uso de la distancia Euclidiana.

$$d = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

donde:

x_i es el i-ésimo valor en el vector x.

y_i es el i-ésimo valor en el vector y.

Número de vecinos

Fijamos un valor para k, habitualmente pequeño, y hacemos que el algoritmo compute una instancia. Fruto de este proceso, el algoritmo selecciona las k instancias datos de entrenamiento más cercanas y se asigna la instancia a la clase más frecuente de entre las k instancias seleccionadas como más cercanas.

Función de ponderación para los vecinos

Si se consideran múltiples vecinos, es razonable que el vecino más cercano (y por lo tanto más similar) tenga una mayor influencia en la predicción. Esto se puede representar mediante una función de peso definida como una función de la distancia del vecino desde el punto de consulta, que devuelve valores más altos para distancias más pequeñas.

Función de predicción

Si se consideran varios vecinos, se necesita un procedimiento para calcular la predicción a partir de las clases o valores objetivo de estos vecinos, ya que pueden diferir y, por tanto, puede que no se obtenga directamente una predicción única.

3.2.2. Gaussian Naive Bayes

Estos clasificadores utilizan la regla de Bayes y algunas hipótesis simplificadoras para modelar las distribuciones de probabilidad condicional de las diferentes clases dados los valores de los atributos descriptivos. Predicen, para cada nueva instancia, la clase más probable clase basada en este modelo de probabilidad. Los clasificadores de Bayes difieren principalmente en el tipo y la estructura de las suposiciones simplificadoras que se hacen [14].

Para obtener un clasificador que prediga la clase más probable, utilizaremos la regla de Bayes.

$$\text{pred(x)} = \arg \quad \max \frac{P(\text{x}|y)P(y)}{P(\text{x})}$$

Donde.

$P(y)$ = es la probabilidad que la hipótesis y sea cierta.

P(x)= es la probabilidad que la hipótesis x sea cierta.

P(x|y)= es la probabilidad de los datos y dado que la hipótesis x era cierta.

3.2.3. Máquinas de Soporte Vectorial

La clasificación basada en situaciones lineales es bastante sencilla, sin embargo, la mayoría de problemas del mundo real representan límites de decisión más complejos. En los últimos años han surgido las llamadas máquinas de soporte vectorial.

Los modelos utilizan clasificadores lineales bien conocidos en combinación con una proyección en espacios de mayor dimensión donde el problema puede ser resuelto o se aproxima bastante. Para esto debemos encontar la función Φ de modo que los datos sean en un nuevo espacio linealmente separables [3]. La Fig. 3.3. Muestra una representación gráfica de los datos en diferentes dimensiones.

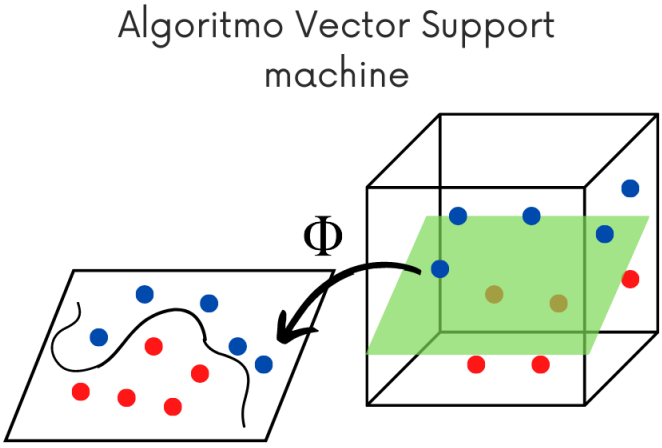


Figura 3.3: Función Φ para proyección de dimensiones.

Si podemos encontrar un hiperplano de separación, podemos observar visualmente lo

que representa nuestro vector $\vec{w}$. No necesitamos utilizar todos nuestros puntos de datos de entrenamiento originales, basta con utilizar los puntos más cercanos a este hiperplano. Estos puntos son los que definen los vectores de soporte. La Fig. 3.4. Muestra los puntos más cercanos, los cuales son los únicos necesarios para nuestro vector de soporte. Se puede observar que hay más de una línea que separa diferentes puntos. Cualquier línea que separe correctamente los puntos de las distintas clases funciona bien. Sin embargo, hay líneas que están más cerca de las instancias de entrenamiento que otras, las cuales usaremos para entrenar el modelo [7]. De esta forma se puede obtener un plano que clasifica las diferentes clases.

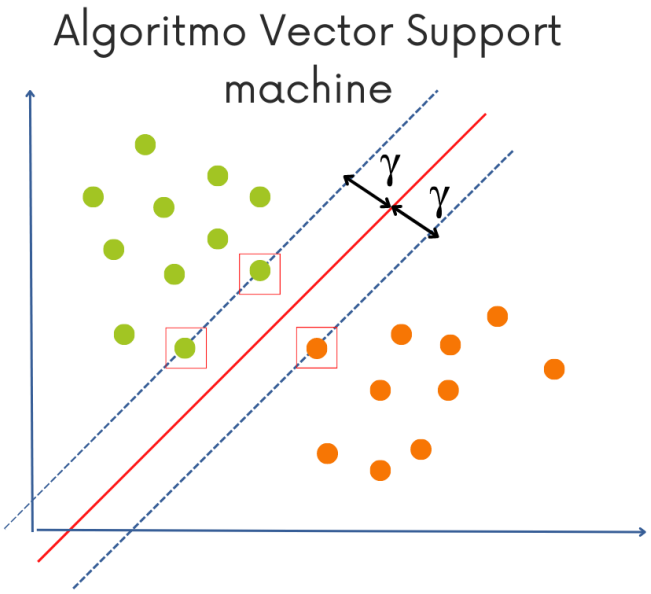


Figura 3.4: El hiperplano que maximiza el margen de error para todas las instancias de entrenamiento.

Este clasificador se mantuvo en las mejores precisiones, siendo en muchos casos el de mayor precisión. Sin embargo, es el algoritmo que más necesita procesamiento y por lo tanto, mayor tiempo para entrenar el modelo. Este clasificador tiene un tiempo hasta 10 veces mayor a los clasificadores K-NN y Bayes .

3.2.4. Redes neuronales, Multi-layer perceptron (MLP)

Las redes neuronales artificiales son una familia de modelos y métodos de entrenamiento que tratan de imitar el aprendizaje de los organismos superiores inspirándose en las redes neuronales biológicas. Las redes neuronales artificiales suelen alcanzar una alta efectividad en tareas de clasificación. Es el algoritmo que modela las neuronas humanas [7]. En la Fig. 3.5. Se puede observar esta comunicación entre neuronas en un cerebro humano.

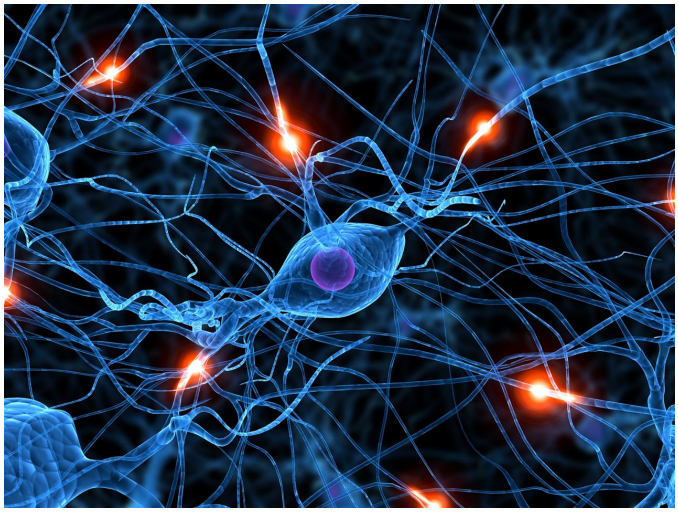


Figura 3.5: Representación de comunicación entre neuronas dentro del cerebro humano. Créditos: Gress.

Un perceptrón multicapa funciona ejecutando las neuronas de las capas de izquierda a derecha. Calculando la salida de cada neurona a partir de sus entradas ponderadas y la función de activación. A este proceso se le conoce como propagación hacia adelante o feed forward. Un perceptrón multicapa se entrena para aplicar una función deseada con la ayuda de un conjunto de datos de casos de muestra, que son pares de valores de entrada y valores de salida asociados.

Los valores de entrada se introducen en el perceptrón multicapa y se calcula su salida. Esta salida se compara con la salida deseada. Si ambos difieren, se ejecuta un proceso denominado propagación para atrás de errores o también conocido como back-propagation que adapta los pesos de conexión y posiblemente otros parámetros de las funciones de activación de forma que se reduzca el error de salida [3].

Para que la retropropagación de errores sea factible, es obligatorio que las funciones de activación sean funciones sigmoidales diferenciables (en forma de s), es decir, la activación de la neurona no debe saltar en el umbral de completamente inactiva a totalmente activa, sino que debe aumentar suavemente. Con este tipo de funciones, la salida de un perceptrón multicapa es una función de las entradas, que también depende de los pesos de conexión y de los parámetros de la función. Como consecuencia, los pesos y los parámetros pueden adaptarse con un esquema de descenso de gradiente, que minimiza la suma de los errores de salida al cuadrado para el conjunto de datos de entrenamiento. Las funciones de activación más utilizadas son la función logística [7].

Para nuestro modelo MLP empleado, cada entrada es un valor de cada canal RGB en cada pixel, que se refiere a los colores que componen la imagen (Red, Green, Blue), como cada imagen tiene un tamaño de 150x150 pixeles y cada uno tiene tres canales, por lo que se tiene un total de 67,500 datos de entrada, posteriormente estas neuronas de entrada están conectadas a la capa oculta por los pesos entre ellos $(\omega_{11}, \omega_{21}, ..., \omega_{T1})$.

La Fig. 3.6. Muestra un diagrama del algoritmo usado para la clasificación de imágenes. Siendo de izquierda a derecha, la capa de entrada donde están las neuronas de entrada, posteriormente la capa oculta con 64 neuronas y por último, nuestra capa de salida con cuatro neuronas.

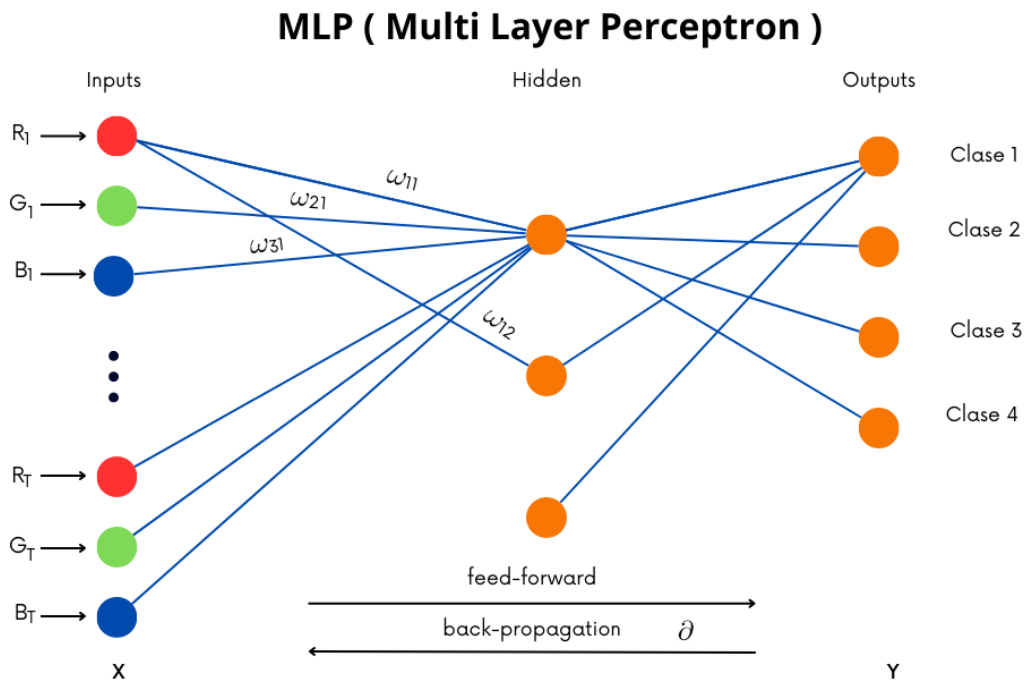


Figura 3.6: Diagrama de algoritmo perceptrón multicapa.

Gradiente descendiente

Cuando la función objetivo es diferenciable, se puede aplicar el método del gradiente. El gradiente es el vector de derivadas parciales con respecto a los parámetros del modelo, apunta en la dirección de mayor ascenso. La idea de la optimización basada en un método de gradiente es comenzar en un punto aleatorio y luego ir un determinado paso en la dirección del gradiente, cuando la función objetivo es máxima, y en dirección opuesta al gradiente cuando la función objetivo sea mínima, que conduce a un nuevo punto en el espacio de los parámetros. Si este punto arroja un mejor valor para la función objetivo, el gradiente en este punto se calcula, y el siguiente punto en la dirección o respectivamente en dirección opuesta del nuevo gradiente. Este procedimiento se continúa hasta que no se consigan más mejoras o se haya realizado un número fijo de pasos de gradiente se ha llevado a cabo [14].

Las imágenes usadas en este trabajo son de tipo RGB, con la combinación de estos tres colores se pueden crear muchísimos colores. Cada pixel contiene un valor de cada canal, este valor puede ir desde el 0 y hasta 255. La Fig. 3.7. Muestra los componentes de una imagen de la galaxia de Antennae, situada a 70 millones de años luz en la constelación de Corvus.

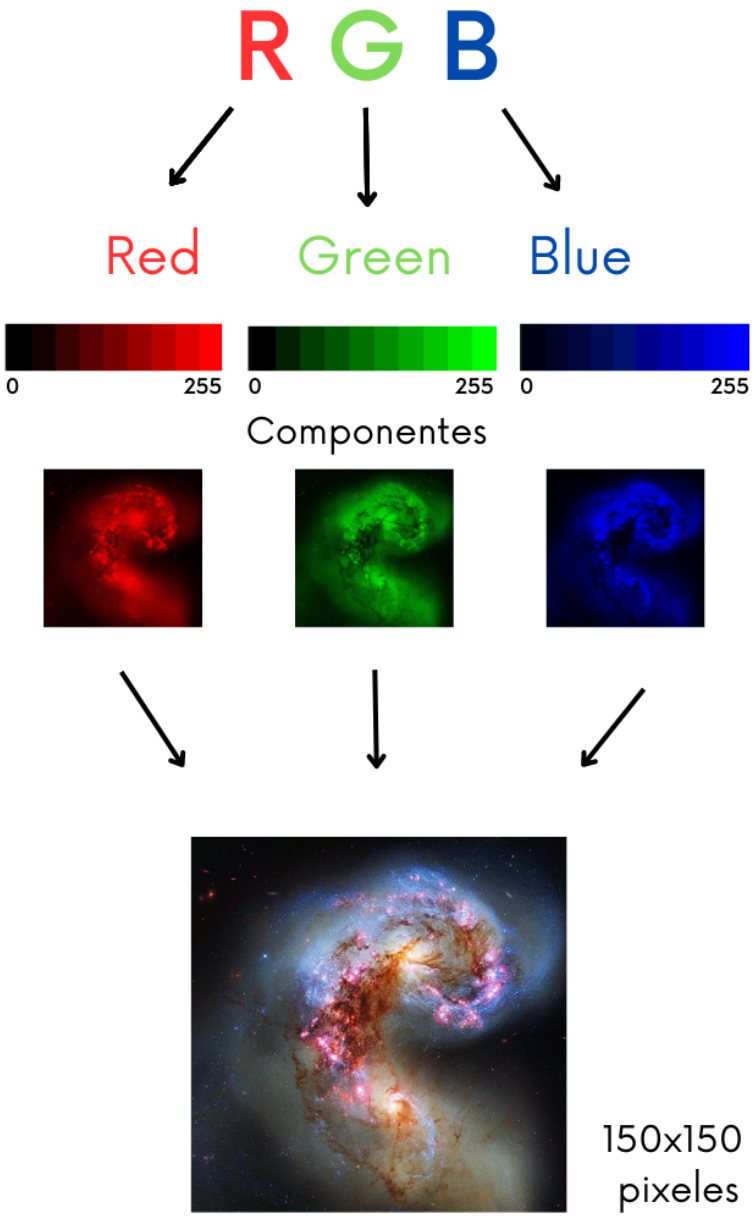


Figura 3.7: Canales RGB que conforman una imagen.

3.3. Validación cruzada

La estimación del error de ajuste de los datos nuevos a partir de un conjunto de datos de prueba que no se han utilizado para aprender el modelo depende de la distribución del conjunto de datos original en datos de entrenamiento y datos de prueba. podríamos tener la suerte de que el conjunto de pruebas contenga más ejemplos fáciles que lleven a una evaluación demasiado optimista del modelo. O podemos tener mala suerte cuando el conjunto de pruebas contiene ejemplos más difíciles y afecte el rendimiento del modelo. Para evitar esta situación se hace el uso de la validación cruzada o mejor conocido como *Cross Validation.* No se basa en una única estimación del error del modelo, sino en varias estimaciones. el conjunto de datos se divide en k subconjuntos de aproximadamente el mismo tamaño. El primero de los k subconjuntos se utiliza como conjunto de prueba, y los otros conjuntos se utilizan como datos de entrenamiento para el modelo. De este modo, obtenemos la primera estimación del error del modelo. A continuación, se repite este procedimiento utilizando cada uno de los otros k subconjuntos como datos de prueba y los subconjuntos restantes como datos de entrenamiento. En total, obtenemos k estimaciones del error del modelo [2]. La media de estos valores se toma como la estimación del error del modelo. Normalmente, se elige $k = 10$.

Capítulo 4

Métodos y Herramientas empleadas

4.1. Python

Python es un lenguaje de programación de alto nivel que se utiliza para desarrollar aplicaciones de todo tipo. Se trata de un lenguaje interpretado, es decir, que no es necesario compilarlo para ejecutar las aplicaciones escritas en Python, sino que se ejecutan directamente por la computadora utilizando un programa denominado interpretador. Python es un lenguaje sencillo de leer y escribir debido a su alta similitud con el lenguaje humano. Además, se trata de un lenguaje multiplataforma de código abierto y, por lo tanto, gratuito, lo que permite desarrollar software sin límites [4]. En la fig. 4.1. Se observa el logo de Python, obtenida de su página oficial (https://www.python.org).

Figura 4.1: Logo de Python.

Con el paso del tiempo, Python ha ido ganando adeptos gracias a su sencillez y a sus amplias posibilidades, sobre todo en los últimos años, ya que facilita trabajar con inteligencia artificial, big data, machine learning y data science, entre muchos otros campos. Python es un lenguaje de programación multiplataforma, algo que permite desarrollar aplicaciones en cualquier sistema operativo con una facilidad asombrosa. Una gran cantidad de tecnologías se llevan muy bien con Python debido a su sencillez y a su gran potencia para el tratamiento de datos [4].

Data analytics y big data: Su simplicidad y su gran número de bibliotecas de procesamiento de datos hacen que Python sea ideal a la hora de analizar y gestionar una gran cantidad de datos en tiempo real.

Data science: La sencillez y la potencia para trabajar con un gran número de datos, unidos al gran número de bibliotecas existentes, hacen que Python sea ideal para este tipo de tareas.

Inteligencia artificial: Su capacidad de plasmar ideas complejas en pocas líneas, unidas al gran número de frameworks existentes, han hecho que Python sea uno de los lenguajes de programación que están impulsando a la IA.

Machine learning: Un entorno de desarrollo integrado es un paquete de software que los desarrolladores utilizan para crear programas. Está destinado a maximizar la productividad mediante la incorporación de componentes estrechamente relacionados con interfaces de usuario simples. Esencialmente, es una herramienta que mejora el proceso de creación, prueba y depuración del código fuente.

4.2. Jupyter Notebook

Para el desarrollo de este proyecto se utilizó el IDE Jupyter Notebook, por su interfaz secilla y productiva. Es una aplicación web de código abierto. Cada desarrollador puede dividir el código en partes y trabajar en ellas sin importar el orden: escribir, probar funciones, cargar un archivo en la memoria y procesar el contenido [6].

Es un entorno de desarrollo interactivo con el live code. Jupyter muestra una ejecución del código a través del navegador web. Si un desarrollador quiere visualizar un gráfico o

una fórmula, escribe el comando deseado en la celda correspondiente. Este enfoque ahorra tiempo y ayuda a evitar errores. La Fig. 4.2. Muestra el logo de Jupyter Notebook, obtenido de su página oficial (https://jupyter.org).

Figura 4.2: Logo de Jupyter notebook.

4.3. Librerías usadas

4.3.1. Pandas

Pandas proporciona estructuras de datos optimizadas y flexibles que se pueden utilizar para manipular datos de serie temporal y datos estructurados, como las tablas y las matrices. Por ejemplo, puede utilizar Pandas para leer, escribir, combinar, filtrar y agrupar datos. Muchas personas lo utilizan para las tareas de ciencia de datos, análisis de datos y ML. Pandas es utilizado para leer la base de datos que contiene el objeto y la clase a la que pertenece. Para este proyecton se usó para leer nuestro documento CSV que funciona como base de datos [4].

4.3.2. Matplotlib

Los desarrolladores utilizan Matplotlib para trazar los datos en gráficos de dos y tres dimensiones (2D y 3D) de alta calidad. Por lo general, se utiliza en las aplicaciones científicas. Con Matplotlib, puede visualizar los datos mostrándolos en diferentes gráficos, como los gráficos de barras y los de líneas. Para este proyecto es utilizado para obtener las matrices de confusión de los clasificadores [4].

4.3.3. Numpy

Proporciona una estructura de datos universal que posibilita el análisis de datos y el intercambio de datos entre distintos algoritmos. Las estructuras de datos que implementa son vectores multidimensionales y matrices con capacidad para gran cantidad de datos. Esta librería es utilizada para crear un arreglo de los datos de entrada y lo puedan leer los algoritmos [7].

4.3.4. Sea Born

Es una librería gráfica basada en matplotlib, especializada en la visualización de datos estadísticos. Se caracteriza por ofrecer un interfaz de alto nivel para crear gráficos estadísticos visualmente atractivos e informativos [4].

4.3.5. SciPy

Proporciona rutinas numéricas eficientes fáciles de usar y opera en las mismas estructuras de datos proporcionadas por NumPy [4].

4.3.6. Scikit-learn

La biblioteca scikit-learn contiene muchas herramientas eficientes para aprendizaje automático y modelado estadístico, incluyendo clasificación, regresión y agrupación. Presenta la compatibilidad con otras librerías como NumPy, SciPy y matplotlib. Ideal para extracción y selección de características de imágenes. Esta es la librería más utilizada, ya que en ella están todos los algoritmos empleados [7].

En la Fig. 4.3. Se muestra el primer cuadro de código, donde se manda a llamar las líbrerias usadas en nuestro programa.

```
In [1]: import numpy as np
        from scipy import ndimage
        from sklearn.model_selection import train_test_split
        import pandas as pd
        from sklearn.neighbors import KNeighborsClassifier
        from sklearn.neural_network import MLPClassifier
        from sklearn.preprocessing import StandardScaler
        from sklearn.svm import SVC
        from sklearn.naive_bayes import GaussianNB
        from sklearn.metrics import confusion_matrix
        import seaborn as sns
        import matplotlib.pyplot as plt
```

Figura 4.3: Librerías usadas en nuestro programa de clasificación.

Puede observar el código completo en el siguiente link:

`https://drive.google.com/drive/folders/1Em7acAavILdFPybetHLkYHrB7sBW07Ga?usp=share_link`

4.4. Base de datos

Para este proyecto se utilizó una base de datos proporcionado por el SDSS. Esta base de datos cuenta con un total de 60,248 imágenes clasificadas en cuatro clases: AGN, Broadline, Starburst y Starforming. Cada imagen tiene un tamaño de 150 x 150 pixeles y con un formato JPG.

Esta base de datos se descargó desde la página oficial del proyecto (classic.sdss.org/dr7). Es una base de datos no equilibrados, ya que tan solo la clase Starforming representa el 70.26 %, la clase Starburst con un 18.75 %, y las clases AGN y Broadline con 5.78 % y 5.21 % respectivamente. Para analizar la evolución de los clasificadores se hicieron sub-bases de datos con diferentes números de imágenes.

Para darse una idea de lo que procesan los algoritmos, mostramos histogramas RGB de las imágenes digitales usadas. Estas muestran la frecuencia que aparece en nuestra imagen

cada valor RGB de los pixeles, este valor puede ir desde el 0 hasta el 255. La Fig. 4.4. Muestra el histograma de una galaxia starburst.

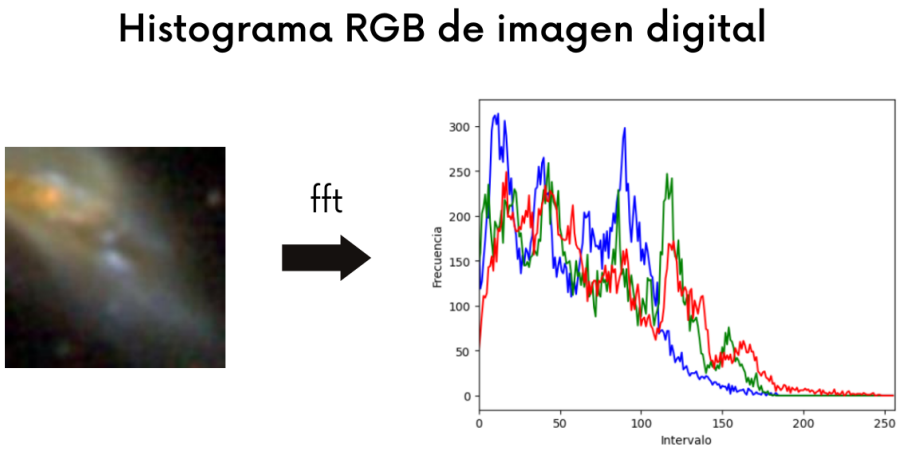


Figura 4.4: Histograma RGB de una galaxia starburst.

Con la información de los histogramas de igual forma se pueden obtener patrones entre las imágenes, aunque con los histogramas es más fácil reconocer la clase de galaxias, hay algunos datos que no parecen corresponder. La Fig. 4.5. Muestra datos de histogramas de diferentes clases.

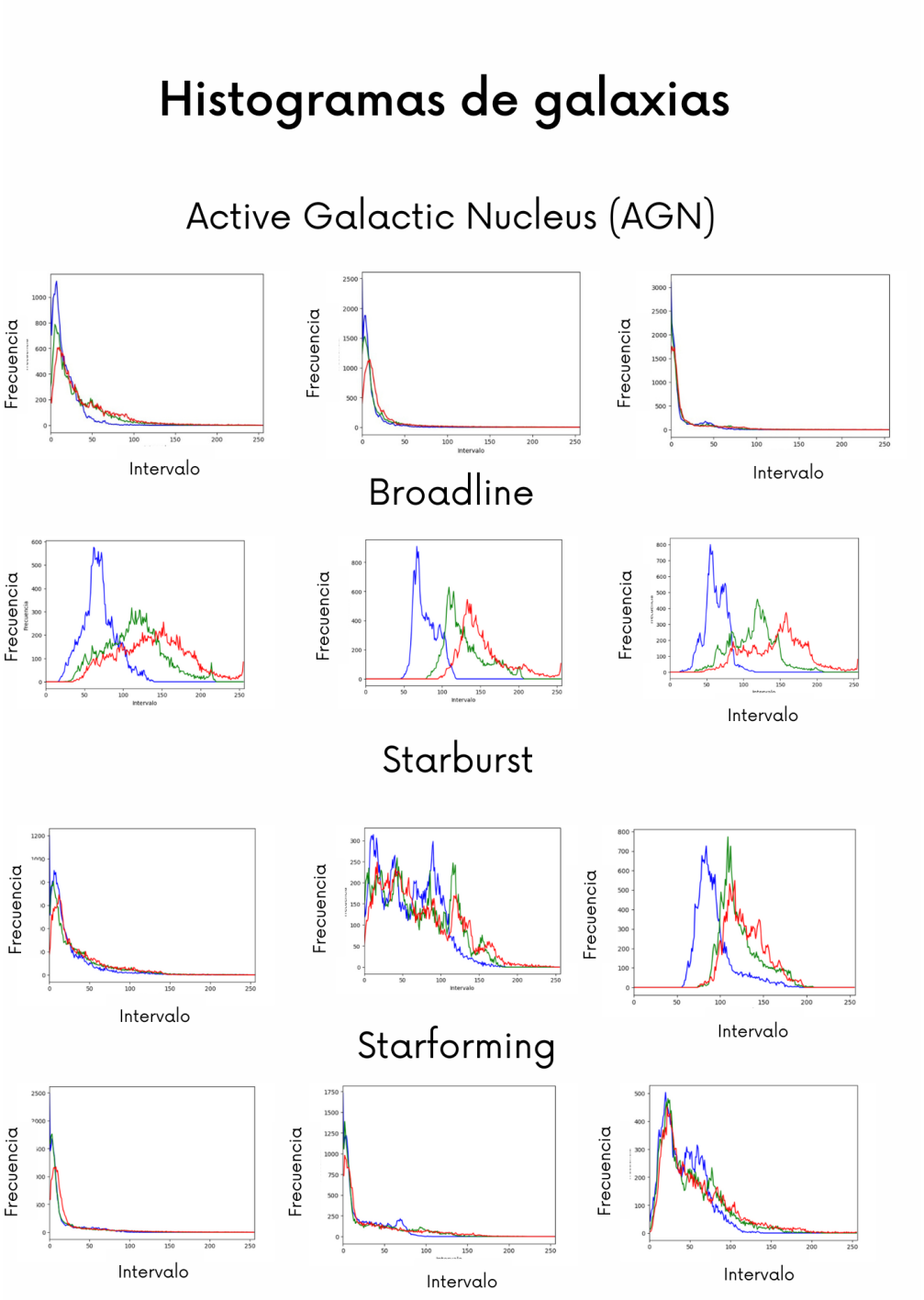


Figura 4.5: Histogramas RGB de galaxias.

Sub-bases de datos

Con datos muy parecidos, nuestros algoritmos tienen ruido en su procesamiento, para reducir este dato se hicieron sub bases de datos con diferente número de clases y diferentes combinaciones. En total se hicieron cuatro clasificadores con tres clases, y seis clasificadores con dos clases, todas con sus respectivas combinaciones posibles, quedando de la siguiente forma:

Cuatro clases

- Broadline
- AGN
- Starforming
- Starburst

Tres clases

- AGN
- Broadline
- Starburst

- AGN
- Broadline
- Starforming

- AGN
- Starburst
- Starforming

- Broadline
- Starburst
- Starforming

Dos clases

- AGN
- Starburst

- AGN
- Broadline

- AGN
- Starforming

- Broadline
- Starburst

- Starburst
- Starforming

- Broadline
- Starforming

Capítulo 5

Resultados

En este capítulo se muestran todos los resultados de los clasificadores. Se divide en dos secciones: clasificadores equilibrados y clasificadores no equilibrados, cada uno de ellos incluye las diferentes combinaciones entre clases. Después de realizar tantas corridas, se notó que los resultados de la precisión es casi la misma aunque se aumente el número de imágenes, pero el tiempo de procesamiento es mucho mayor mientras se aumentan los datos, por lo que se decidió usar un total de 1,000 imágenes para los casos de clasificadores de tres y cuatro clases. Se realizaron 28 clasificadores, lo que da un total de 280 corridas.

En cuestión de procesamiento computacional y tiempo que tardaba cada clasificador, se obtuvieron tiempos completamente diferentes. Para clasificadores con hasta máximo 1,000 datos de entrada, el tiempo en terminar el programa fue de aproximadamente 8 minutos, siendo menor mientras los datos eran cada vez menos. Sin embargo, los clasificadores de 3,000 tomaron un aproximado de 30 minutos, los clasificadores de 5,000 datos, tomaron casi 90 minutos. Los clasificadores con 10,000 datos tomaban más de 6 horas y los clasificadores con 20,000 datos tomaron poco más de 20 horas. Todos estos datos se procesaron con equipos de cómputo ordinarios, con memorias ram de 16 GB.

Se toma el cross validation de cada uno de los clasificadores. Se comienza con 4 clases (AGN, Starburst, starforming, broadline) con diferente número de imágenes, de igual forma, el número de datos por galaxias no es proporcional.

5.1. Clasificadores no equilibrados

A continuación se muestran los resultados del cross validation de los clasificadores para cuatro clases con un dataset no equilibrado, por lo que el número de imágenes va en aumento para analizar su comportamiento. En la columna de clases se muestra los tipos de galaxias que se procesaron, igualmente se muestra en paréntesis el porcentaje de cada clase en el dataset. Después se muestra la precisión de cada algoritmo (K-NN, Naive Bayes, Máquinas de soporte vectorial, Perceptrón multi capa.)

Resultados de clasificadores 4 clases

		% Precisión			
datos	Clases (% en dataset)	K-NN	Naive Bayes	SVM	MLP
250	• Broadline (2.8) • AGN (7.6) • Starforming (71.6) • Starburst (18.0)	68.0	23.2	67.7	57.2
500	• Broadline (6.2) • AGN (3.0) • Starforming (74.6) • Starburst (16.2)	73.5	21.4	72.6	61.3
750	• Broadline (2.93) • AGN (5.73) • Starforming (75.34) • Starburst (16.0)	71.46	56.66	74.52	65.81
1,000	• Broadline (3.5) • AGN (5.6) • Starforming (74.5) • Starburst (16.4)	72.1	58.45	73.4	68.0

		% Precisión			
datos	Clases (% en dataset)	K-NN	Naive Bayes	SVM	MLP
2,000	• Broadline (4.6) • AGN (5.55) • Starforming (72.8) • Starburst (17.05)	71.65	59.02	73.19	66.75
3,000	• Broadline (5.0) • AGN (6.14) • Starforming (71.2) • Starburst (17.7)	70.24	55.05	72.97	65.76
4,000	• Broadline (4.95) • AGN (6.425) • Starforming (71.75) • Starburst (16.875)	70.50	55.37	71.50	61.0
5,000	• Broadline (5.22) • AGN (6.54) • Starforming (71.68) • Starburst (16.56)	70.97	58.95	71.95	65.97
10,000	• Broadline (4.96) • AGN (6.38) •Starforming (70.96) •Starburst (17.7)	72.26	60.35	71.96	65.90
20,000	• Broadline (4.80) • AGN (6.06) • Starforming (70.32) •Starburst (18.81)	72.57	52.25	71.35	67.43

De estos resultados se puede observar la evolución de los datos según se aumentan la cantidad de datos. Estos clasificadores con datos no equilibrados de cuatro clases son los que tienen el mayor número de datos en todo este trabajo, el clasificador con más datos contiene un total de 20,000 imágenes.

En la parte de la precisión de los clasificadores, el algoritmo Máquinas de soporte vectorial es el que dió resultados más altos, seguido de k-Vecinos más cercanos, en tercer lugar los perceptrones multi capa y por último los clasificadores de Gaussian naive bayes.

La Fig. 5.1. Muestra la gráfica de la evolución de la precisión con diferentes números de datos. Se puede observar que en un principio los clasificadores aumentan su precisión con el aumento de datos, sin embargo, aproximadamente despues de 1,000 datos, los clasificadores se comportan de manera constante y en algunos casos la precisión disminuye. Por lo que los resultados de 1,000 datos es muy aproximado a los resultados obtenidos con 20,000 datos.

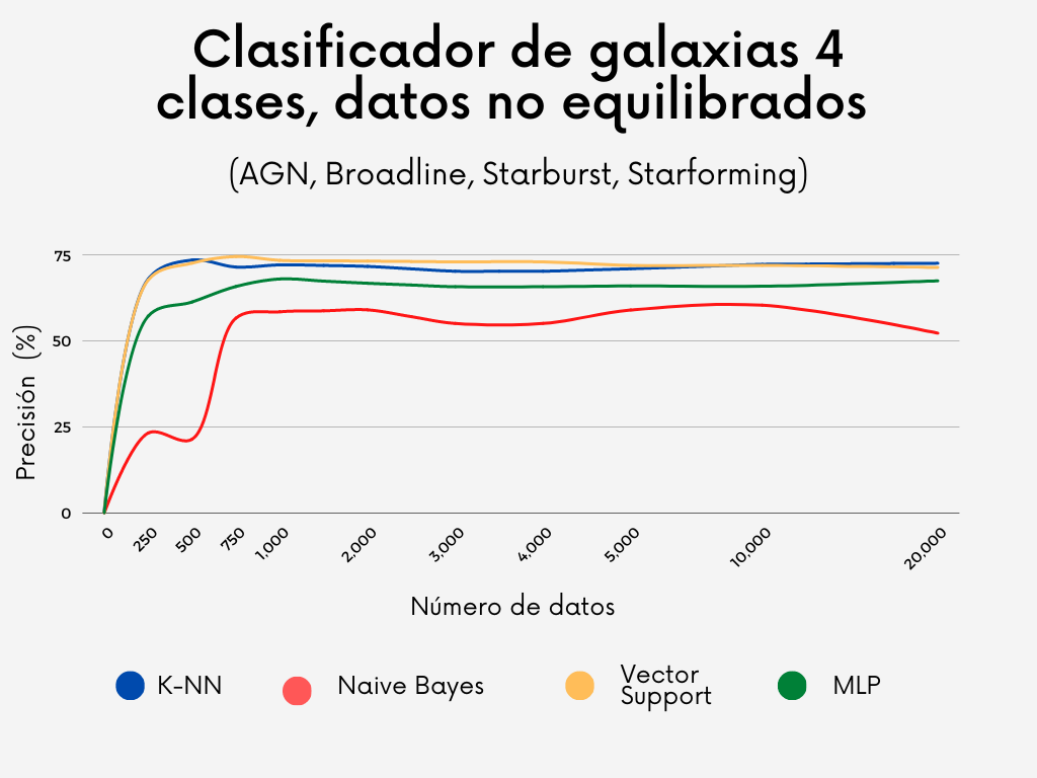


Figura 5.1: Evolución de precisión con diferente número de datos.

Con esta información se decidió hacer diferentes clasificadores pero con solo 1,000 datos, ya que se obtienen resultados bastante similares a clasificadores con muchos más datos.

En nuestra base de datos podemos encontrar imágenes casi idénticas por su forma, pero son de clases totalmente diferente, por este motivo se decidió hacer clasificadores con menos clases y así observar el comportamiento de los clasificadores con solo tres clases. Para estos clasificadores solo se uso un total de 1,000 datos y se hicieron todas las combinaciones posibles con tres clases.

A continuación se muestran los resultados de las combinaciones posibles con clasificadores de tres clases. Aunque el número de datos es constante, el porcentaje de la base de datos que ocupa cada clase no es equilibrado.

Resultados de clasificadores 3 clases

		% Precisión			
datos	Clases (% en dataset)	K-NN	Naive Bayes	SVM	MLP
1,000	• AGN (5.8) • Starforming (77.1) • Starburst (17.1)	75.45	54.15	75.75	69.35
1,000	• Broadline (15.0) • AGN (10.0) • Starforming (75.0)	77.05	65.50	79.05	73.0
1,000	• Broadline (15.0) • AGN (10.0) • Starburst (75.0)	78.15	63.75	84.40	78.10
1,000	• Broadline (15.0) • Starforming (75.0) • Starburst (10.0)	75.2	54.35	75.2	70.0

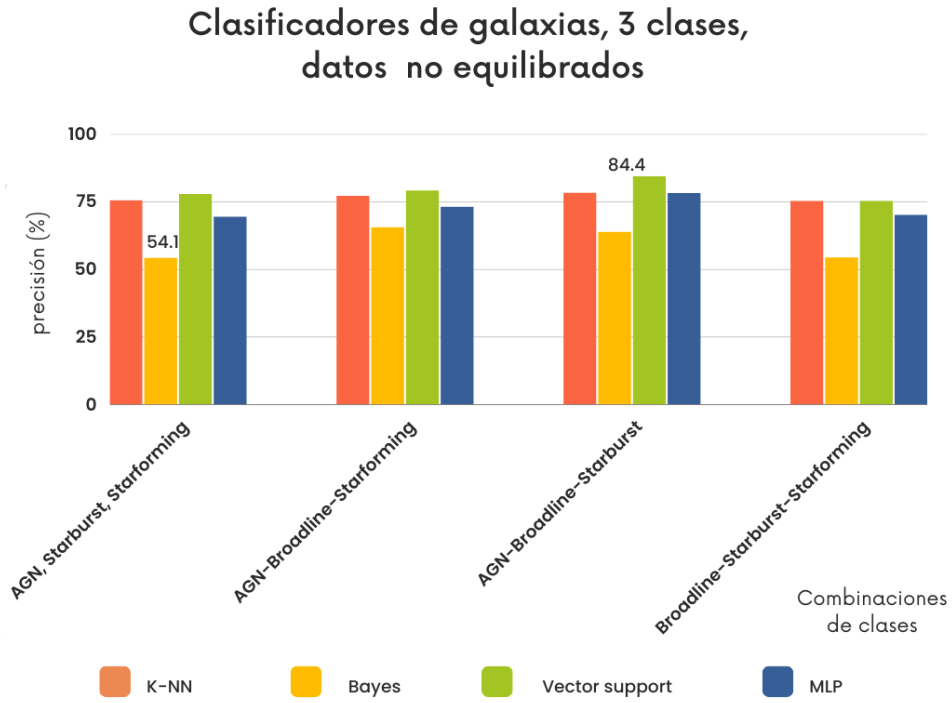


Figura 5.2: Resultados de combinaciones de clasificadores no equilibrados.

5.2. Clasificadores equilibrados

Después de observar el comportamiento de los clasificadores no equilibrados, se hizo la siguiente pregunta, si una sola clase ocupa más de tres cuartas partes del total de los datos, hay poca información sobre las demás clases ¿con datos equilibrados se obtendrán mejores resultados?

Por lo que se hicieron bases de datos donde todas las clases tuvieran el mismo número de imágenes y ver el comportamiento de la precisión de los clasificadores.

A continuación se muestra la tabla de resultados de clasificadores con datos equilibrados y un máximo de 3,000 imágenes.

Resultados de clasificadores 4 clases

		% Precisión			
datos	Clases (% en dataset)	K-NN	Naive Bayes	SVM	MLP
250	• Broadline (25.2) • AGN (24.8) • Starforming (25.2) • Starburst (24.8)	45.20	32.20	51.60	43.20
500	• Broadline (25) • AGN (25) • Starforming (25) • Starburst (25)	47.30	31.80	54.0	52.50
750	• Broadline (24.93) • AGN (25.06) • Starforming (24.93) • Starburst (25.06)	42.92	32.93	53.59	45.46
1,000	• Broadline (25) • AGN (25) • Starforming (25) • Starburst (25)	45.81	33.23	52.01	45.21

		% Precisión			
datos	Clases (% en dataset)	K-NN	Naive Bayes	SVM	MLP
2,000	• Broadline (25) • AGN (25) • Starforming (25) • Starburst (25)	47.40	34.37	52.62	47.10
3,000	• Broadline (25) • AGN (25) • Starforming (25) • Starburst (25)	50.16	33.66	56.16	48.83

La Fig. 5.3. Muestra el comportamiento de los clasificadores con datos equilibrados, mostrando el mismo patrón que tienen los clasificadores con datos no equilibrados.

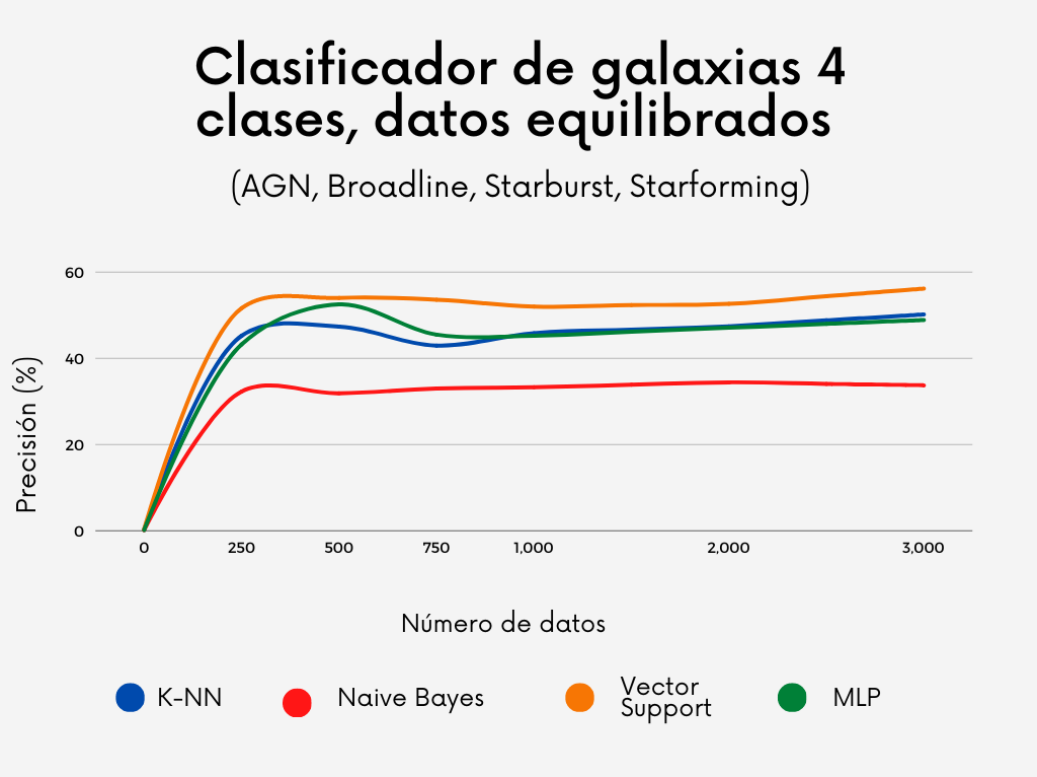


Figura 5.3: Evolución de precisión con diferente número de datos.

Resultados 3 clases

		% Precisión			
datos	Clases (% en dataset)	K-NN	Naive Bayes	SVM	MLP
1,000	• AGN (33.3) • Broadline (33.3) • Starburst (33.4)	58.8	46.45	66.10	61.35
1,000	• Broadline (33.3) • Starburst (33.3) • Starforming (33.4)	60.75	45.3	69.0	64.3
1,000	• AGN (33.3) • Broadline (33.3) • Starforming (33.4)	52.30	44.40	59.35	53.45
1,000	• AGN (33.3) • Starburst (33.3) • Starforming (33.4)	51.20	37.50	58.80	54.15
	Resultados 2 clases				
1,000	• AGN (50) • Starburst (50)	76.70	61.90	82.85	81.15
1,000	• AGN (50) • Starforming (50)	64.75	53.70	69.15	65.15
1,000	• Broadline (50) • AGN (50)	64.55	59.55	65.90	59.85
1,000	• Starburst (50) • Broadline (50)	83.1	63.7	87.2	85.7
1,000	• Starburst (50) • Starforming (50)	74.10	62.15	80.70	79.0
1,000	• Broadline (50) • Starforming (50)	79.25	61.20	79.60	78.55

En la Fig. 5.4. Se observa una gráfica de barras con los resultados de precisión, se muestran cuatro diferentes combinaciones entre clases. Siendo el clasificador Broadline, Starburst y Starforming, con una precisión de 69.

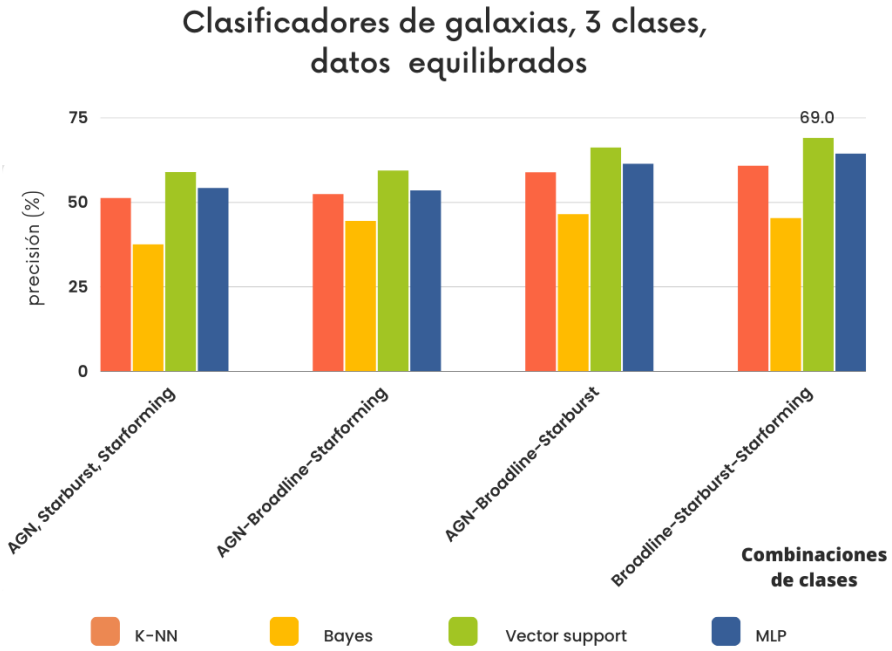
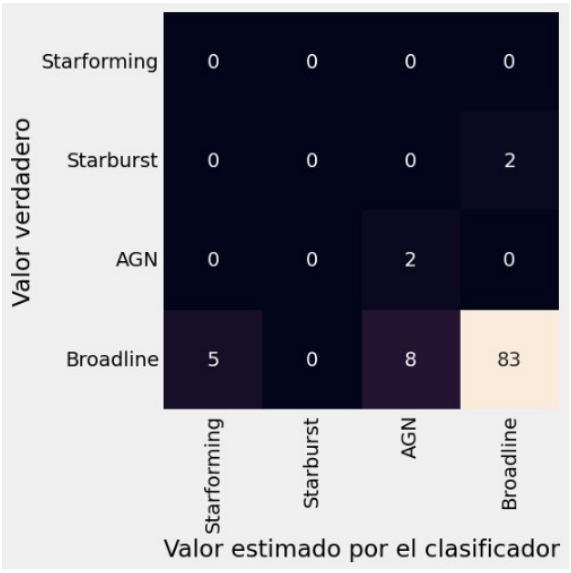


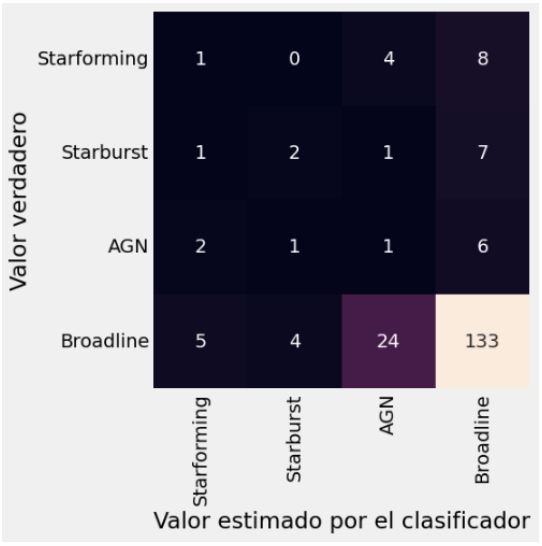
Figura 5.4: Resultados de combinaciones de clasificadores con datos equilibrados.

Una buena herramienta para el análisis de datos, es la matriz de confusión la cual permite analizar los resultados del trabajo de un algoritmo de aprendizaje supervisado. En esta matriz se presenta en forma de tabla, de manera que en cada columna aparece el número de predicciones de cada clase, mientras que cada fila muestra el número real de instancias de cada clase. A continuación se muestran las matrices de confusión de las mejores corridas.

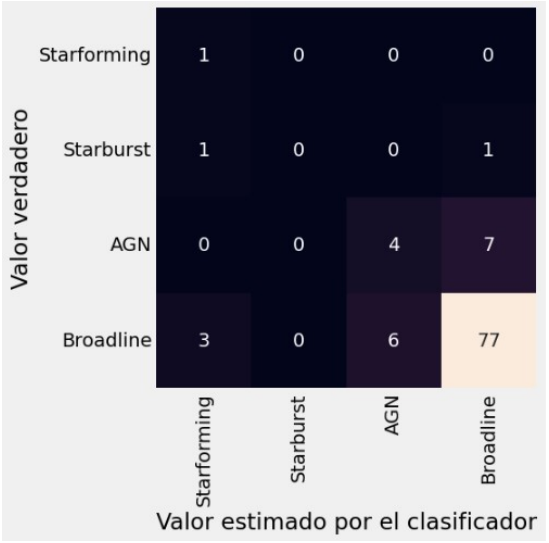
Resultados matrices de confusión para corridas individuales en clasificadores de cuatro clases con datos no equilibrados.



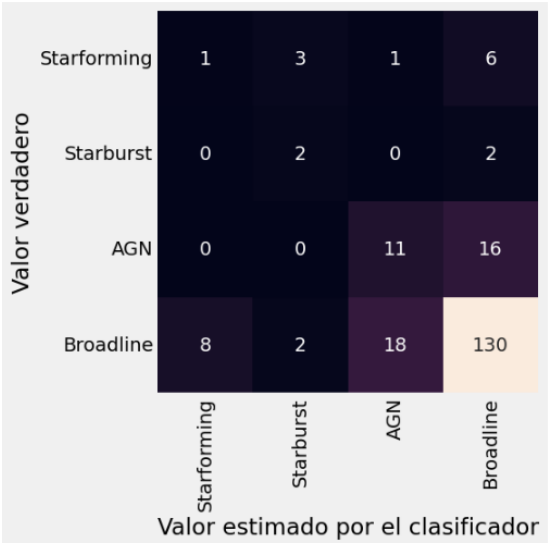
(a) Algoritmo k-Vecinos más cercanos en clasificador con 500 datos, precisión= 85



(b) Algoritmo Naive Bayes en clasificador con 750 datos, precisión= 70.6



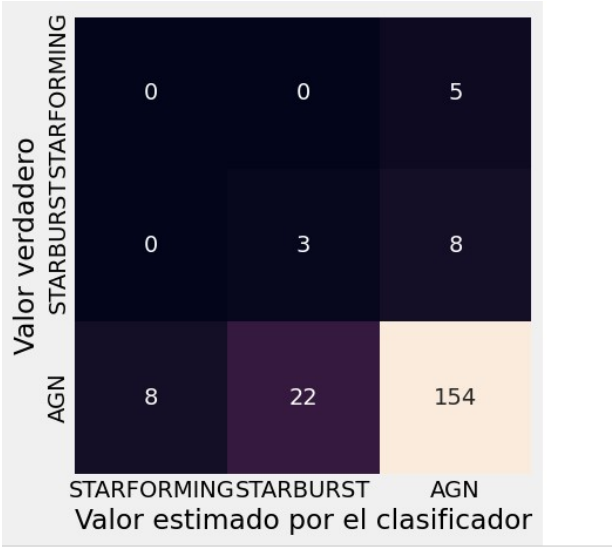
(c) Algoritmo Máquinas de soporte vectorial en clasificador con 500 datos, precisión= 82



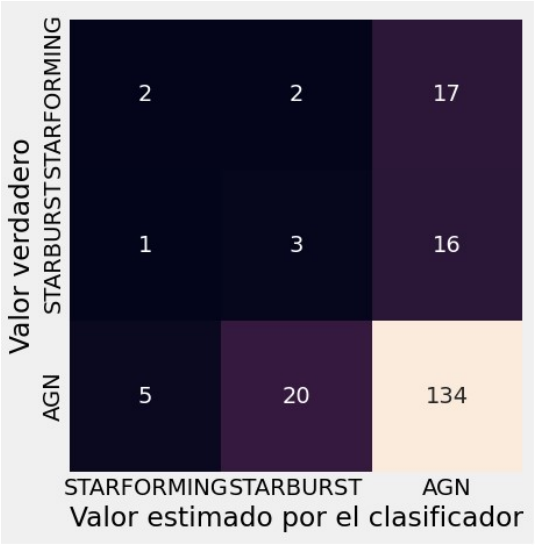
(d) Algoritmo Perceptrón multicapa en clasificador con 1,000 datos, precisión= 72

Figura 5.5: Resultados matrices cuatro clases, no equilibrado.

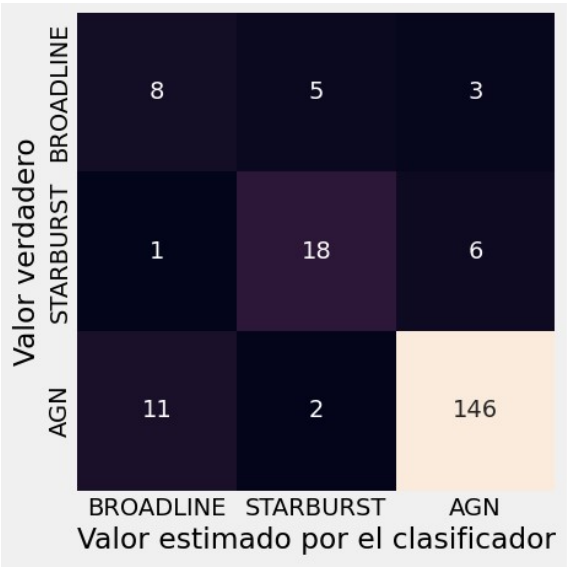
Resultados matrices de confusión para corridas individuales en clasificadores de tres clases, todos con 1,000 datos no equilibrados.



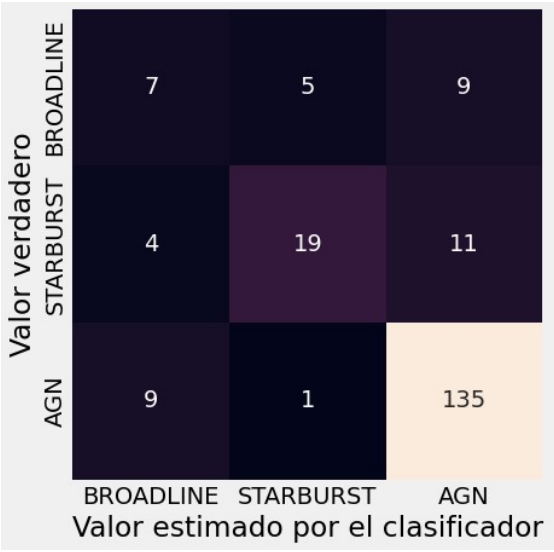
(a) Algoritmo k-Vecinos más cercanos, precisión= 78.5



(b) Algoritmo Naive Bayes precisión= 69.5



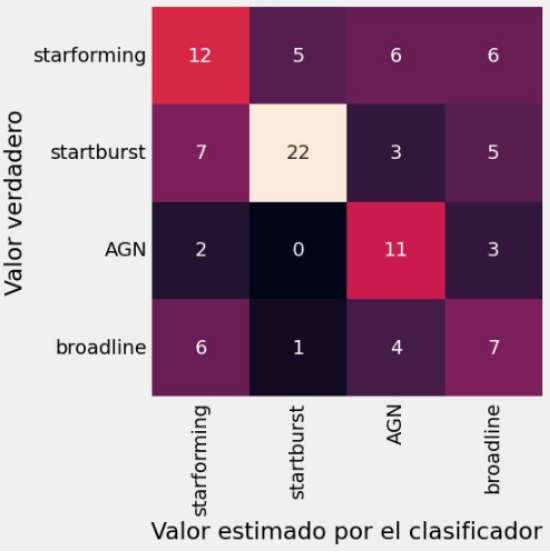
(c) Algoritmo Máquinas de soporte vectorial, precisión= 86



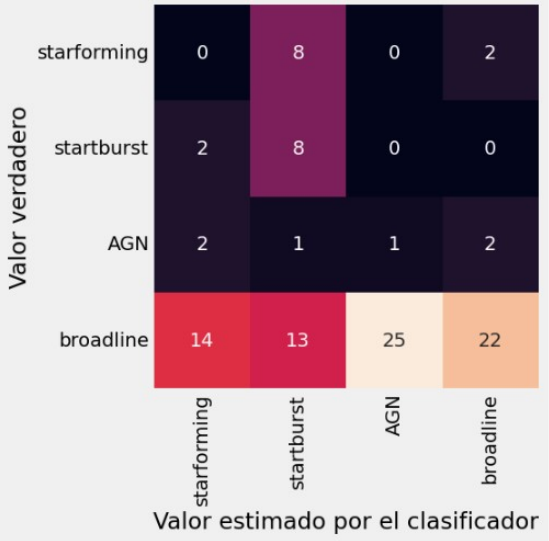
(d) Algoritmo Perceptrón multicapa precisión= 80.5

Figura 5.6: Resultados de matrices, tres clases y 1,000 datos.

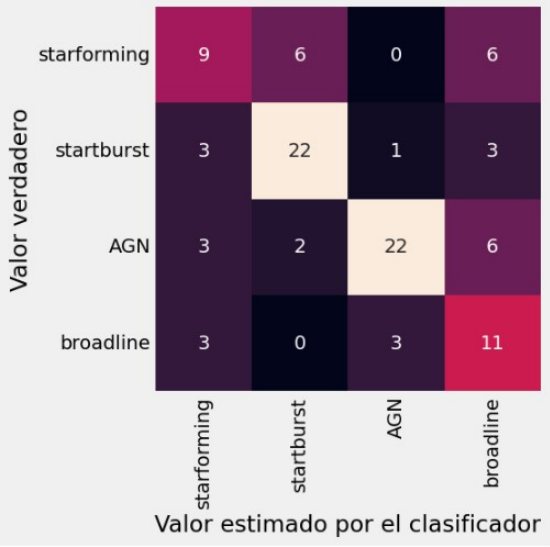
Resultados matrices de confusión para corridas individuales en clasificadores de cuatro clases con datos equilibrados, todos pertenecientes al clasificador con 500 datos.



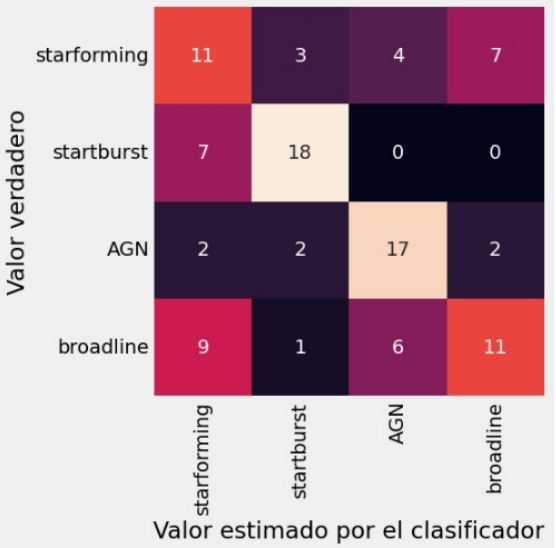
(a) k-Vecinos más cercanos, precisión= 52



(b) Naive Bayes, precisión= 31



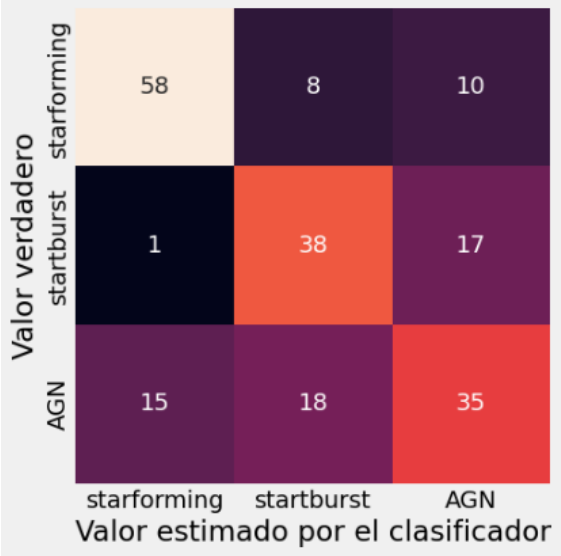
(c) Máquinas de soporte vectorial, precisión= 64



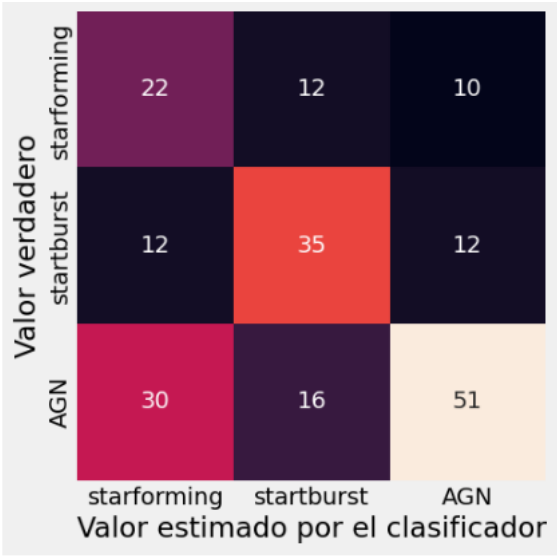
(d) Perceptrón multicapa, precisión= 57

Figura 5.7: Resultados de matrices , cuatro clases, datos equilibrados.

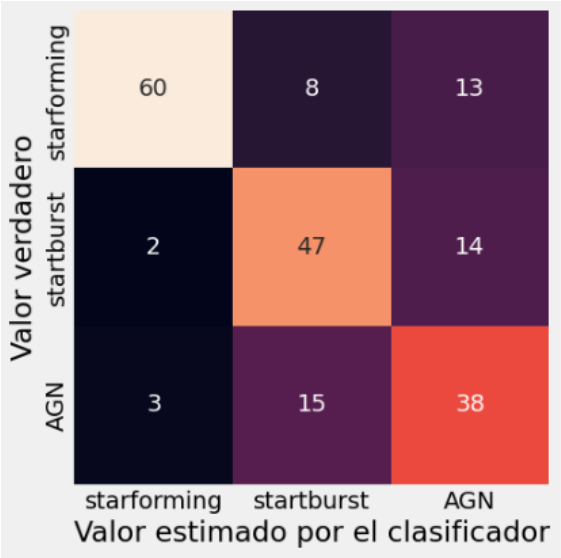
Resultados matrices de confusión para corridas individuales en clasificadores de tres clases, todos con 1,000 datos equilibrados, todas las matrices pertenecen al clasificador AGN, Starburst, Starforming.



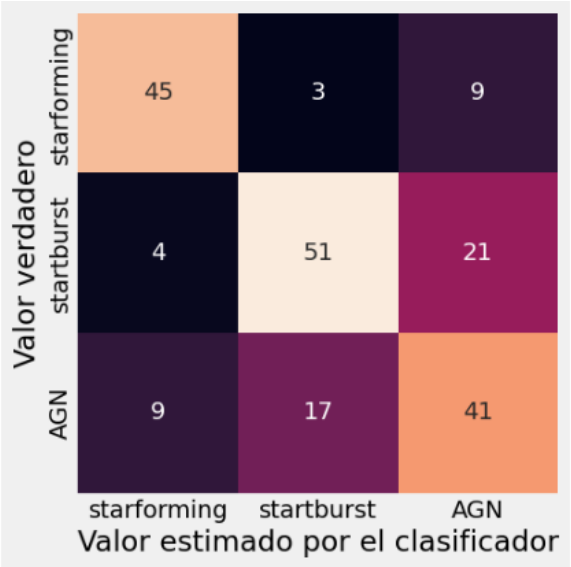
(a) Algoritmo k-Vecinos más cercanos, precisión= 67



(b) Algoritmo Naive Bayes. precisión= 54



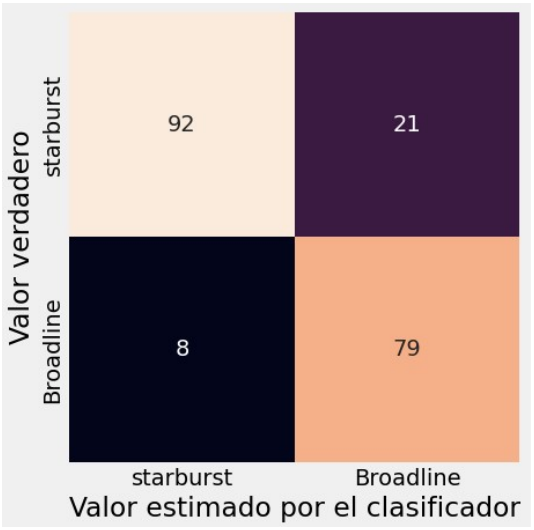
(c) Algoritmo Máquinas de soporte vectorial, precisión= 73



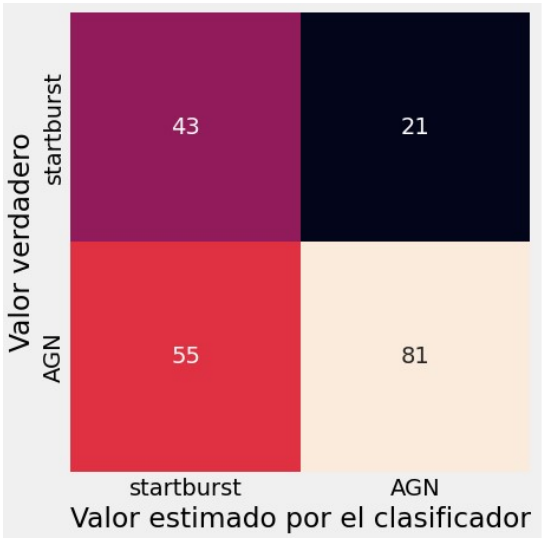
(d) Algoritmo Perceptrón multicapa, precisión=68.5

Figura 5.8: Resultados de matrices, tres clases, datos equilibrados.

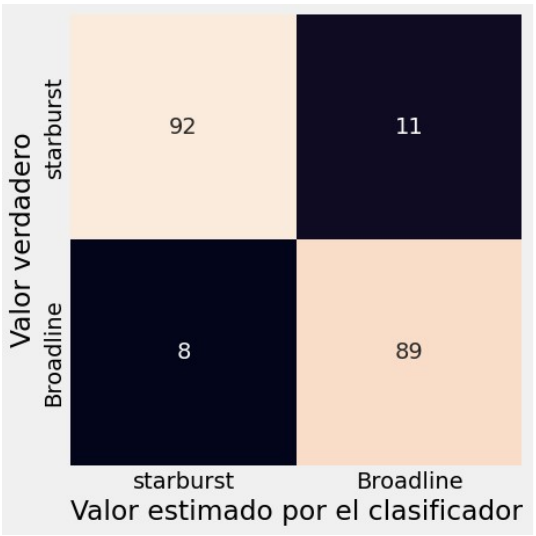
Resultados matrices de confusión para corridas individuales en clasificadores de dos clases con 1,000 datos equilibrados.



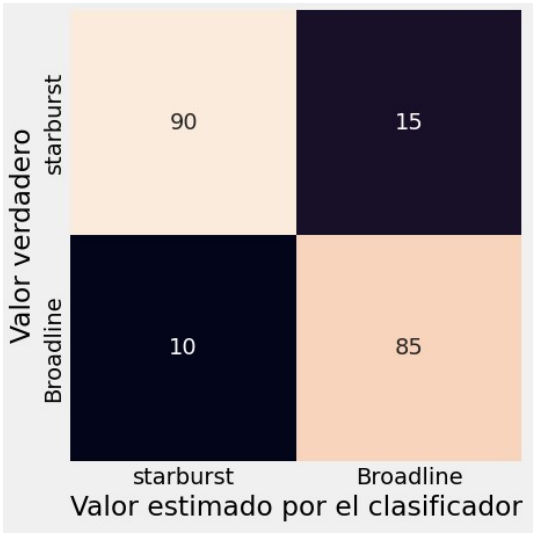
(a) Algoritmo k-Vecinos más cercanos, precisión=85



(b) Algoritmo Naive Bayes precisión=62



(c) Algoritmo Máquinas de soporte vectorial, precisión =90



(d) Algoritmo Perceptrón multicapa, precisión= 90

Figura 5.9: Resultados de matrices, dos clases, datos equilibrados.

Capítulo 6

Conclusiones y trabajo a futuro

6.1. Conclusiones

Se analizaron los resultados de clasificadores con diferente número de datos, los más pequeños con 250 imágenes y el más grande con 20,000 imágenes. Se realizaron clasificadores con cuatro, tres y dos clases, con todas sus combinaciones entre clases. De igual forma, los clasificadores se hicieron con datos equilibrados y no equilibrados. Estos resultados y gráficas se estudiaron, con esto podemos concluir que: Todos los clasificadores convergen a un resultado muy similar en sus corridas con diferentes números de datos (ver Fig. 4.1). Los algoritmos dan resultados similares sin importar si son 1,000 ó 20,000 imágenes. En los clasificadores con datos no equilibrados se obtuvo mayor precisión que los clasificadores con datos equilibrados. Sin embargo, en los clasificadores con datos equilibrados, el algoritmo de perceptrón multicapa aumenta significativamente su precisión, mientras los demás algoritmos disminuyen su precisión. De igual forma se nota que entre menos clases haya, los clasificadores obtienen mayor precisión. Los clasificadores de dos clases obtuvieron los resultados más altos, siendo el mas alto con 87.2 % y clases starburst y broadline con el algoritmo máquinas de soporte vectorial. Con las sub-bases de datos, se observan colores característicos de cada clase y en algunos casos, formas que distingue cada clase. La clase Starburst es la más fácil de identificar visualmente, de igual forma las combinaciones con esta clase dieron de los mejores resultados en la precisión. Hay casos de en las que los patrones de cada clase son completamente diferentes aunque sean de la misma clase,

por lo que realizar un procesamiento previo de datos podría aumentar la precisión de los clasificadores.

6.2. Trabajo a futuro

Se pretende continuar este trabajo con algoritmos mucho más potentes y sofisticados, como lo son las redes neuronales convolucionales. De igual forma, realizar clasificadores con mayor número de datos (hasta 60,000) para analizar el comportamiento de los algoritmos con el aumento de datos. Es de interés tratar de hacer clasificadores entre diferentes objetos astronómicos, hay bases de datos mucho más grandes en el SDSS (180 y 240 mil datos), para clasificaciones entre galaxias, quasars y estrellas. Se busca mejorar la precisión de los clasificadores, por lo que es necesario hacer pruebas con bases de datos que contengan pre-procesamiento de imágenes que contienen patrones visualmente para analizar los resultados. Hay algoritmos que muestran mejora en ciertas clases, por lo que hacer ensambles entre algoritmos y ver su comportamiento, es un tema interesante.

Bibliografía

[1] Boselli A. *A panchromatic view of galaxies*. Editorial WILEY-VCH, 2012.

[2] Borgelt C. y Berthold M. *Guide to Intelligent Data Analysis*. Editorial Springer, 2010.

[3] Bigus P. y Bigus J. *Constructing Intelligent Agents Using Java*. Editorial WILEY, 2001.

[4] Hill C. *Learning Scientific Programming with Python*. Editorial Cambridge University Press, 2020.

[5] Bladley w. y Dale A. *An Introduction to Modern Astrophysics*. Editorial Springer, 2014.

[6] Torres J. *Python Deep Learning*. Editorial Alfaomega, 2017.

[7] Heaton J. *Artificial Intelligent for Humans. Deep Learning and neuronal networks*. Editorial Heaton Research, Inc, 2015.

[8] Karttunen H. y Kroger P. *Fundamental Astronomy*. editorial Springer, 2004.

[9] Beckman V. y Shrader C. *Active Galactic Nuclei*. Editorial WILEY-VCH, 2012.

[10] Thomas T. *Explorations, An introduction to astronomy*. Mosby- year book Inc, 1996.

[11] «Morphological Classification of Galaxies by Shapelet Descomposition in the Sloan Digital sky Survey». En: (2014). Ed. por Brandon C. y Timothy A. URL: `https://iopscience.iop.org/article/10.1086/380934`.

[12] «Galaxy Formation and Evolution.» En: (2006). Ed. por Mo H. y Van F. URL: `https://typeset.io/papers/galaxy-formation-and-evolution-5d3ajhb2lp`.

[13] «Machine learning technique for morphological classification of galaxies from the SDSS.» En: (2021). Ed. por Vasylenko M. Vavilova B. Dobrycheva D. URL: `https://arxiv.org/abs/2209.12194`.

[14] Sugomori Y. *Java Deep Learning Essentials*. Editorial Packt Publishing, 2016.