Uso de herramientas de Machine Learning para la clasificación de una muestra de galaxias activas a partir de observaciones fotométricas y espectroscópicas

Katherine Andrea Caballero Soto

Facultad de Ciencias
Observatorio Astronómico Nacional
Bogotá, Colombia
2024

Uso de herramientas de Machine Learning para la clasificación de una muestra de galaxias activas a partir de observaciones fotométricas y espectroscópicas

Katherine Andrea Caballero Soto

Tesis presentada como requisito para optar por el título de:
Magíster en Ciencias, Astronomía

Director(a):
Ph.D., Mario Armando Higuera Garzón
Profesor Titular
Facultad de Ciencias
Universidad Nacional de Colombia

Línea de investigación:
Núcleos Activos de Galaxias (AGNs)
Grupo de investigación:
SAGAN

Universidad Nacional de Colombia
Facultad de Ciencias
Observatorio Astronómico Nacional
2024

Dedicatoria.

Dedico esta tesis a mi familia por su apoyo incondicional y haber creído en mí, a mi abuela Mélida de Soto, a mi madre Liliana Soto, a mi tía Martha Soto y mi hermano Julián Soto. A mi novio Jesús Barriga por haber estado a mi lado cada día, de inicio a fin en este sueño y darme la fuerza para lograr finalizarlo.

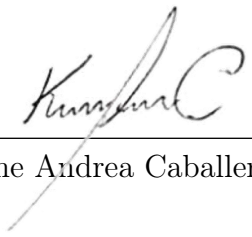
''El cosmos es todo lo que es, todo lo que fue y todo lo que será. Nuestras más ligeras contemplaciones del cosmos nos hacen estremecer: sentimos como un cosquilleo nos llena los nervios, una voz muda, una ligera sensación como de un recuerdo lejano o como si cayéramos desde gran altura. Sabemos que nos aproximamos al más grande de los misterios.''

Carl Sagan

Declaración

Me permito afirmar que he realizado ésta tesis de manera autónoma y con la única ayuda de los medios permitidos y no diferentes a los mencionados en el presente texto. Todos los pasajes que se han tomado de manera textual o figurativa de textos publicados y no publicados, los he reconocido en el presente trabajo. Ninguna parte del presente trabajo se ha empleado en ningún otro tipo de tesis.

Bogotá., 19/01/2024



Katherine Andrea Caballero Soto

Agradecimientos

En este trabajo de investigación agradezco al Observatorio Astronómico Nacional por ser el espacio donde recibí todas las enseñanzas necesarias para llevar a cabo este proyecto.

A mi director, el profesor Mario Armando Higuera por su apoyo constante, enseñanzas, dedicación y liderazgo en todas las etapas de este trabajo.

Al profesor Eduard Alexis Larrañaga por el aprendizaje adquirido en programación, área fundamental para llevar a cabo el desarrollo del proyecto.

Un especial agradecimiento a los profesores José Gregorio Portilla, Santiago Vargas, Giovanni Pinzón quienes me abrieron las puertas y me impulsaron a realizar esta Maestría.

Finalmente a todas las personas cercanas que estuvieron durante este proceso, mi más sincero agradecimiento.

Resumen

Uso de herramientas de Machine Learning para la clasificación de una muestra de galaxias activas a partir de observaciones fotométricas y espectroscópicas

La adquisición de datos astronómicos ha experimentado una revolución, tanto en calidad como en complejidad, durante las últimas décadas, por lo tanto, es necesario no solo desarrollar nuevos métodos para procesar y analizar grandes volúmenes de datos, sino también asegurar que las técnicas aplicadas para extraer la información de los datos sea lo más óptima posible.

Los núcleos galácticos activos son fuentes astrofísicas energéticas impulsadas por acreción de material en agujeros negros supermasivos en las galaxias y presentan huellas observacionales únicas que cubren el espectro electromagnético. Las clasificaciones de estos objetos están relacionadas con las diferencias intrínsecas del AGN y reflejan principalmente variaciones en un número de parámetros astrofísicos.

Este trabajo pretende discriminar y encontrar los mejores datos disponibles tomados de la base de datos del Sloan Digital Sky Survey para explorar la utilidad de una serie de parámetros que permitan establecer un sistema de clasificación más completo, a partir de una variedad de diferentes conjuntos de líneas de emisión que dan información sobre las condiciones del gas ionizado. Se hace uso de líneas de baja probabilidad, además de incluir datos fotométricos asociados a filtros de color, esto con el fin de obtener variables de entrada necesarias para un algoritmo clasificatorio desarrollado con Machine Learning y, con el cual, se obtienen los diferentes diagramas diagnóstico, capaces de catalogar una muestra de galaxias activas que se encuentran en el Universo.

Una vez explorados los distintos modelos de aprendizaje, se evalúa la eficiencia de cada uno, y se determina que el aprendizaje profundo, también conocido como Deep Learning, alcanza una precisión de aproximadamente el 99 %, cuando los parámetros que componen la red neuronal sean optimizados. Esto posibilita la visualización de la distribución poblacional según su actividad energética dominante.

Palabras clave: AGN; Aprendizaje Automático; Aprendizaje Profundo; Diagramas Diagnóstico; Espectroscopía; Fotometría

Abstract

Using Machine Learning tools for the classification of a sample of active galaxies based on photometric and spectroscopic observations

The acquisition of astronomical data has experienced a revolution in both quality and complexity over the last decades. Therefore, it is necessary not only to develop new methods for processing and analyzing large volumes of data but also to ensure that the techniques applied to extract information from the data are as optimal as possible.

Active Galactic Nuclei (AGN) are energetic astrophysical sources powered by the accretion of material onto supermassive black holes in galaxies. They exhibit unique observational signatures that span the electromagnetic spectrum. The classifications of these objects are tied to the intrinsic differences of the AGN, primarily reflecting variations in numerous astrophysical parameters.

This study aims to discriminate and identify the best available data from the Sloan Digital Sky Survey database. The objective is to explore the utility of a range of parameters that could establish a more comprehensive classification system. This system will be based on a variety of different emission line sets, which will provide information about the conditions of the ionized gas. We will utilize low-probability lines and include photometric data associated with color filters. This approach aims to obtain the necessary input variables for a classification algorithm developed with Machine Learning. This algorithm will yield diagnostic diagrams capable of cataloging a sample of active galaxies found in the Universe.

Once the different learning models have been explored, the efficiency of each one is evaluated, and it is determined that Deep Learning reaches an accuracy of approximately 99 %, when the parameters that make up the neural network are optimized. This makes it possible to visualize the population distribution according to its dominant energy activity.

Keywords: AGN; Deep Learning; Diagnostic Diagram; Machine Learning; Photometry; Spectroscopy

Lista de figuras

Contenido

1 Objetivos

1.1 Objetivo General

Desarrollar un algoritmo en lenguaje máquina, "Machine Learning", que permita construir diagramas diagnóstico de clasificación para la distinción de la actividad energética dominante en una muestra de galaxias activas.

1.2 Objetivos Específicos

1. Realizar la búsqueda de emisiones fotométricas y espectroscópicas de una muestra de galaxias utilizando bases de datos públicas del Sloan Digital Sky Survey (SDSS), Infrared Astronomical Satellite (IRAS), NASA/IPAC Extragalactic Database (NED), SIMBAD Astronomical Database - CDS, con el fin de construir un catálogo de datos suficientemente amplio en número y observaciones, condición necesaria para la ejecución eficiente de un algoritmo en Machine Learning.

2. Construir los criterios de clasificación para galaxias con actividad energética activa, basado entre las razones de líneas espectrales de emisión de Hα, Hβ, [SII], [NII], [OI], [OII], [OIII] y filtros fotométricos.

3. Desarrollar un algoritmo de aprendizaje y clasificación, mediante códigos de programación utilizando herramientas de Machine Learning, para construir diagramas diagnóstico equivalentes al tradicional BPT que cataloguen grupos de galaxias activas.

4. Contrastar la discriminación obtenida del algoritmo de Machine Learning con los resultados en la literatura.

2 Estado Del Arte

2.1 Núcleos Activos de Galaxias

En la región central de algunas galaxias se observa una prominente emisión de energía que cubre varios órdenes de magnitud, esta emisión se extiende ampliamente a través del espectro electromagnético, con un pico en los rayos UV-Óptico, en las bandas de rayos X, infrarrojos y radio [*Padovani et al.*. 2017]. Además de ello, se reflejan características adicionales, tales como variabilidad de la emisión, existencia de jets relativistas emisores de radio y espectros de energía en el óptico asociados a nubes fotoionizadas. Dichos objetos se denominan Núcleos Activos de Galaxias (por sus siglas en inglés Active Galactic Nuclei, AGNs) [*Osterbrock*. 1989]. Los agujeros negros supermasivos son la parte esencial en la dinámica de un AGN, ubicados en el centro de una galaxia, considerados como una pieza clave en la evolución y formación estelar galáctica. Los agujeros negros con masas mayores o iguales a $10^6 M_\odot$, se encuentran rodeados de nubes que están siendo sometidas a radiaciones fotoionizantes que surgen del disco de acreción [*Peterson*. 1997]. Los núcleos galácticos activos involucran las fuentes de luminosidad más poderosas del Universo, van desde galaxias cercanas que emiten alrededor de 10^{40}erg/s, hasta cuásares distantes que emiten más de 10^{47}erg/s [*Andrew*. 1999].

2.1.1 Clasificación de AGNs

En el universo, los AGNs pueden ser divididos en dos tipos según su actividad nuclear, radio silenciosos (radio-quiet) o radio ruidosos (radio-loud). A continuación se detalla cada uno de ellos:

Radio Silenciosos

- **Seyfert:** AGNs de alta luminosidad, cuya emisión abarca desde radio hasta los rayos X, y en los que la galaxia anfitriona es una galaxia espiral. La energía emitida por el

núcleo en el óptico es comparable a la que emite toda la galaxia anfitriona ($\approx 10^{11} L_{\odot}$). Su espectro se caracteriza por poseer líneas de emisión intensas debido a las transiciones de los gases ionizados [*Nwankpa et al.*. 2020].

Las galaxias Seyfert, a su vez, se subdividen en dos tipos, Sy1 y Sy2, atendiendo a sus características espectrales.

- Seyfert I: Presentan altas densidades electrónicas ($n_e \approx 10^9 \mathrm{cm}^{-3}$), líneas de emisión permitidas anchas H I, He I y He II, con una medida de la velocidad del ancho de la línea a la mitad de la altura $v_{\mathrm{FWHM}} \approx 10^4 \mathrm{km/s}$, además de líneas prohibidas asociadas a gases ionizados de baja densidad ($n_e \approx 10^2 - 10^3 \mathrm{cm}^{-3}$), para líneas tales como [OIII] 4959Å, 5007Å, [NII] 6548Å, 6584Å y [SII] 6716Å, 6731Å, con un v_{FWHM} entre 10^2 km/s y 10^3 km/s [*Higuera*. 2012].

- Seyfert II: Presentan gases ionizados de baja densidad con espectros de líneas angostas y $v_{\mathrm{FWHM}} \approx 10^2$ y 10^3 km/s. Su espectro muestra líneas anchas, tanto prohibidas como permitidas, con los mismos anchos de las líneas angostas de las galaxias tipo Seyfert 1, generando así un continuo menos intenso [*Cutiva Alvarez*. 2018].

- **LINER:** Núcleos activos de baja luminosidad, caracterizados por un espectro de emisión de líneas de baja ionización, angostas e intensas en [OI] 6300Å, [NII] 6548Å, y bandas de absorción estelar, representando un estado intermedio entre las galaxias activas y no activas. Las galaxias tipo LINER (por sus siglas en inglés Low Ionization Narrow Emission-line Regions) pueden distinguirse de las Seyfert basándose en los cocientes de flujo, donde los valores de [OIII] 5007Å/Hβ con respecto a [NII] 6583Å/Hα son más bajos, en comparación con las regiones H II, donde sus valores son mayores [*Osterbrock*. 1989]. Su luminosidad en Hα oscila entre 10^{37}-$10^{41}\mathrm{erg/s}^{-1}$, y constituyen 2/3 de la población de AGNs.

- **QSO:** Los cuásares comprenden la subclase más luminosa de AGN, caracterizados por tener un núcleo muy brillante no resuelto, con luminosidades ópticas mayores que las de las galaxias que las albergan, cuya emisión abarca desde longitudes de onda de radio hasta los rayos X. La expresión QSO es un acrónimo de "Quasi-Stellar Object", ya que tienen una apariencia estelar (fuente puntual) en imágenes ópticas con tamaños angulares inferiores a 7". Los espectros de los cuásares son similares a los de las galaxias Seyfert, exceptuando que las bandas de absorción estelar son muy débiles, además de que sus componentes de líneas anchas de emisión, descritos por una ley de potencias, son mucho más intensas que las estrechas. Este tipo de galaxias activas se clasifican como lo objetos más distantes que se encuentran en el inicio de su etapa evolutiva, con valores altos de redshift ($z \approx 8$). [*Peterson*. 1997]

Radio Ruidosos

Los núcleos galácticos activos radio-ruidosos (RLAGNs) son un conjunto de la población de galaxias activas que desempeñan un papel importante en la evolución de las mismas, regulando la formación de estrellas en galaxias masivas [*Hardcastle et al.*. 2019]. La acumulación de materia en el agujero negro supermasivo central (SMBH) puede impulsar chorros relativistas de partículas cargadas (electrones, positrones, y/o protones) y campo magnético, que emergen desde una distancia cercana al radio de Schwarzschild, produciendo emisiones de sincrotrón, la fuente poderosa de la emisión en radio [*Urry*. 2003].

- **Radio Galaxias:** Tipos de AGNs donde las fuentes de radio más intensas suelen centrarse en el núcleo, lóbulos y los jets alejados del objeto central. Se encuentran, generalmente, en galaxias tipo elípticas gigantes o quásares [*Cutiva Alvarez*. 2018]. Luminosas en frecuencias de radio (10 MHz - 100 GHz), sus espectros ópticos evidencian líneas de emisión anchas (BLRGs, Broad-line Radio Galaxies) y líneas estrechas (NLRGs, Narrow-line Radio Galaxies).

- **Blázares:** Se definen como un sub-grupo de los cuásares cuya característica principal es que los jets o chorros que emergen de ellos están ubicados en dirección a la Tierra, y se pueden clasificar en dos tipos: Las galaxias BL Lacertae (BL Lac) son objetos en los que sus líneas de emisión espectrales son difíciles de detectar o están completamente ausentes. Su emisión es dominada por los chorros relativistas de partículas que se mueven a velocidades cercanas a la velocidad de la luz, y se detectan a redshift bajo. Los blazares Variables Ópticamente Violentos (OVV), con líneas anchas más intensas y brillantes, permiten ser detectados a un redshift más alto [*Cutiva Alvarez*. 2018].

2.2 Modelo Unificado de los AGNs

La diversidad de objetos que presentan actividad nuclear, observados mediante varios instrumentos, así como las diferencias en sus espectros de emisión, condujeron inicialmente a su clasificación en diferentes categorías. No obstante, hacia finales de la década de 1980, se llegó a la conclusión de que todos comparten atributos comunes y que las variaciones en los espectros de líneas están vinculadas a las distintas orientaciones desde las cuales se observan. A raíz de estas investigaciones emergió un modelo integrador desarrollado por Miller & Antonucci en 1983, que proporciona una explicación para las disparidades observadas en los espectros de varios tipos de galaxias activas [*Antonucci*. 1993]. Estos modelos explicaron la ausencia de líneas anchas en los espectros de los AGNs de tipo 2 como resultado de efectos de la orientación desde la línea de visión. Este principio también se aplicaría a los objetos de tipo 1, pero con diferentes ángulos de observación.

La **Figura 2-1** es la representación gráfica del modelo unificado, donde se muestra cada una de las partes de las que se compone un AGN, como se describe a continuación.

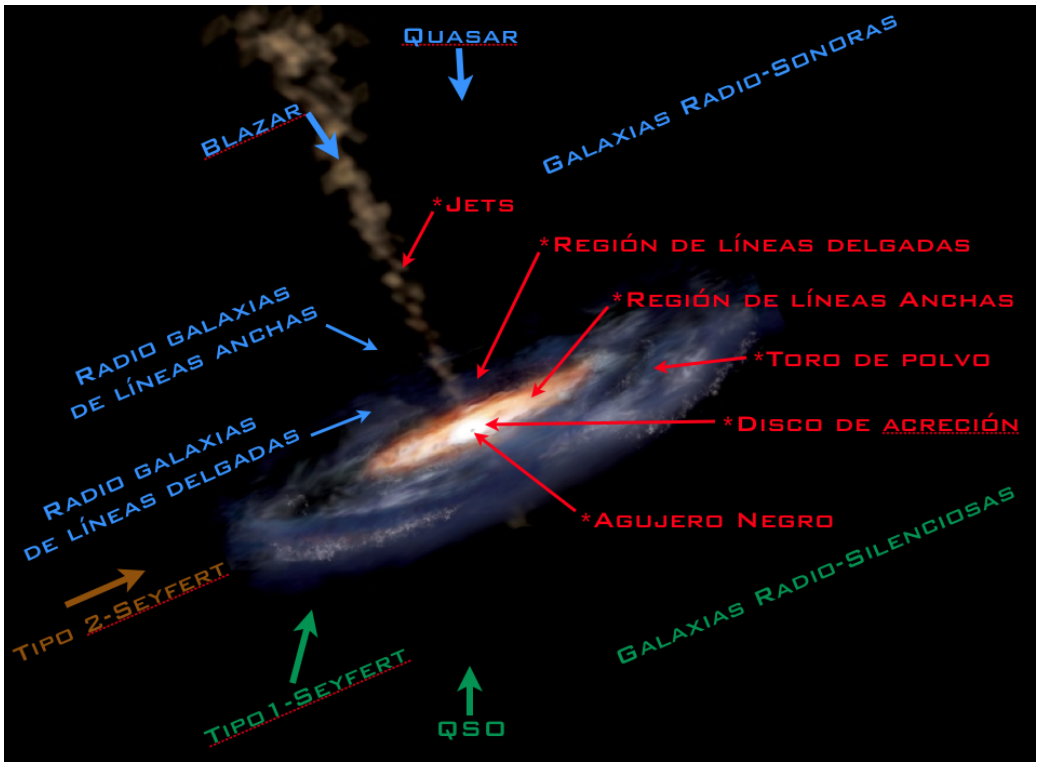


Figura 2-1: Modelo Unificado de AGNs. Adaptado por: [*Higuera*. 2012]

2.2.1 Agujero Negro Supermasivo

La gran cantidad de producción de energía en un AGN se asocia a la existencia un agujero negro supermasivo, el cual consume el gas en caída gravitacional proveniente del disco de acreción que lo rodea. La masa de la fuente central se obtiene utilizando el criterio del teorema del virial $M \approx v^2 r/G$. La velocidad de dispersión del gas, v, se obtiene del ancho de las líneas de emisión, mientras que la distancia r al centro de las nubes emisoras se obtiene de la determinación de los tiempos de reverberación [*Higuera*. 2012]. Los valores de la masa central equivalen a valores entre $10^6 M_\odot$ y $10^{10} M_\odot$.

2.2.2 Disco de Acreción

El disco de acreción se sitúa a una distancia cercana al centro, el cual constituye la fuente de material que se acumula hacia el Agujero Negro Supermasivo, siendo este proceso una de las principales fuentes de energía para el Núcleo Activo de Galaxia. El disco puede

alcanzar temperaturas de hasta 10^5 K debido a la fricción viscosa entre las nubes de gas, transformando la energía potencial gravitacional en radiación de diversas longitudes de onda, que va desde rayos X hasta ondas de radio. Debido a la alta conductividad del material, en el disco se genera un campo magnético variable, cerca de la superficie del mismo, induciendo un campo eléctrico significativo que acelera las partículas hacia afuera, siguiendo las líneas del campo magnético [*Rueda Vargas*. 2020]. Estas partículas alcanzan velocidades cercanas a la velocidad de la luz, dando lugar a la radiación de sincrotrón. La dinámica de estos objetos se rige por la conservación del momento angular y se extiende en el plano del núcleo activo. La masa del disco se estima en alrededor de $10^8 M_\odot$, con un pico de emisión de fotones con una energía de alrededor de 100 eV [*Peterson*. 1997].

2.2.3 Región de Líneas Anchas

La región de líneas anchas (BLR) es aquella parte del AGN que se extiende después del disco de acreción a una distancia menor a 1 parsec, mientras que el borde externo se encuentra limitando con la región del toro y la región de líneas delgadas. Dentro de sus características está la presencia de líneas con anchos equivalentes a velocidades entre 1000 y 25000 km/s y temperaturas de 10^4 K [*Osterbrock*. 1989]. La densidad de electrones de esta zona es de $n_e \approx 10^9 - 10^{11} \mathrm{cm}^{-3}$. Las líneas de emisión del espectro corresponden a la serie de Balmer Hα (6563Å), Hβ (4861Å), Hγ (4341Å), la serie de Lyman Lyα (1216Å), y las líneas de los iones Mg II (2798Å), [CIII] (1909Å) y CIV (1549Å) [*Peterson*. 1997].

2.2.4 Región de Líneas Angostas

Las nubes de la región de líneas angostas (NLR) se encuentran a una distancia mayor de la fuente central, su tamaño puede estar comprendido entre 100 parsecs hasta 1000 parsecs, con temperaturas que oscilan entre los 10000 K a los 25000 K y valores de velocidad del gas que van de los 200 - 1000 km/s [*Osterbrock*. 1989]. La densidad de electrones en la región de líneas delgadas es lo suficientemente baja ($n_e \approx 10^2 - 10^4 \mathrm{cm}^{-3}$) como para permitir transiciones prohibidas de los átomos, esto genera en el espectro de emisión varias líneas como [OII] (3727Å), [NeIII] (3869Å), [OIII] (5007Å), [OI] (6300Å), [NII] (6584Å), [SII] (6716Å) [*Peterson*. 1997].

2.2.5 Región Toroidal

El aspecto más destacado del modelo unificado es el toro de polvo ópticamente espeso, ubicado entre la BLR y la NLR, que oculta la visión de la BLR y el mecanismo central, esta región es una fuente emisora de continuo y de líneas en el rango del infrarrojo medio.

La idea inicial de una región de polvo que rodea los AGNs proviene del trabajo de Donald Osterbrock quien, en 1978 sugirió, que las galaxias Seyfert 1 y 2 son, en esencia, las mismas Seyfert y que la existencia del toroide provee un obscurecimiento anisotrópico de la emisión central [*Osterbrock*. 1989].

2.3 Trazadores de actividad galáctica

Las emisiones que provienen de las galaxias, ya sea por formación estelar o por un núcleo activo, se extienden a lo largo del espectro electromagnético, siendo dominantes en algunas longitudes de onda. Es importante entender que las galaxias con fuerte formación estelar, llamadas Starburst [*Heckman*. 2005], suelen tener un comportamiento similar a las tipo AGN, en cuanto a que la mayor parte de sus emisiones están en el infrarrojo, con altas luminosidades, sin embargo, en otras regiones del espectro difieren en sus características. Es por ello que definir claros criterios de identificación es fundamental para catalogar el tipo de galaxia que se está analizando, y, para ello, se ha desarrollado, desde la década de los 80s, diferentes técnicas.

2.3.1 Diagramas Diagnóstico BPT

La clasificación de galaxias activas se realiza a través del análisis de espectros, es por ello que en 1981, Baldwin, Philips y Terlevich fueron los pioneros en presentar una serie de diagramas de diagnóstico que involucraban razones de flujo de líneas de emisión, con ionización dominante, para así separar y clasificar galaxias por su tipo de actividad energética, sea esta asociada a un núcleo activo o a la presencia de una fuerte tasa de formación estelar. Se estudiaron distintos objetos extragalácticos utilizando espectros con líneas de emisión, para poder clasificarlos de acuerdo a su mecanismo de excitación, como la fotoionización por estrellas O y B, la excitación por una ley de potencia de una fuente continua, ondas de choque o por la excitación por nebulosa planetaria [*Olave Rojas*. 2014].

Una primera aproximación se muestra en la **Figura 2-2**, utilizando los cocientes de intensidades de ([OII]λ3726Å/[OIII]λ5007Å) vs. ([OIII]λ5007Å/H$\beta\lambda$4861Å),([NII]λ6584Å/H$\alpha\lambda$6563Å) vs. ([OIII]λ5007Å/H$\beta\lambda$4861Å), donde se visualiza la distinción y agrupación de los diferentes objetos en varias zonas del diagrama. Las galaxias bajo la curva representan la presencia de formación estelar, mientras que aquellas que están sobre la curva se asocian a un núcleo galáctico activo [*Baldwin et al.*. 1981]. Las razones de línea fueron seleccionadas de modo que las líneas están muy próximas en longitud de onda, para así evitar problemas por efectos de extinción.

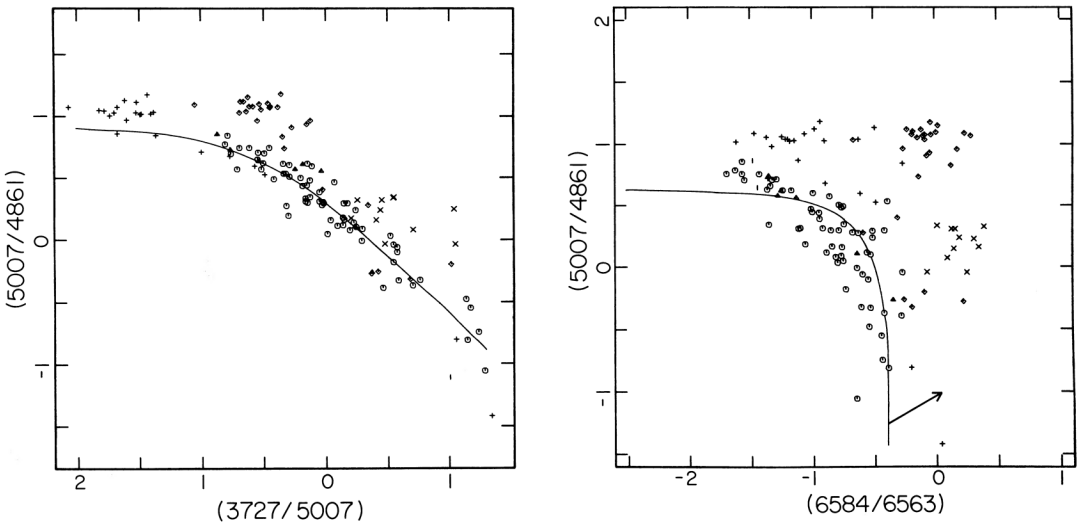


Figura 2-2: Diagramas Diagnóstico BPT para las diferentes razones de línea, expresadas en logaritmo, con correcciones de enrojecimiento. Los círculos representan regiones HII normales, los triángulos corresponden a regiones HII extragalácticas, los diamantes son galaxias con núcleo activo, las cruces(+) nebulosas planetarias y las equis(x) corresponde a galaxias con calentamiento radiativo por choques. Imagen extraída de [*Baldwin et al.*. 1981]

Años más tarde, Veilleux y Osterbrock (1987) derivaron una clasificación para distinguir el mecanismo de ionización del gas nebular con mayor precisión, donde se adaptaron líneas divisorias para separar las galaxias tipo Starburst de las de núcleo activo, como se observa en la **Figura 2-3**, para diferentes líneas de emisión, como lo son [SII], [NII] y [OI].

La relación [OIII]/Hβ es un indicador del nivel de ionización y temperatura, mientras que las relaciones [OI]/Hα y [SII]/Hα son indicaciones de la importancia relativa de una zona parcialmente ionizada producida por fotoionización de alta energía [*Veilleux & Osterbrock*. 1987]. Los autores en esta investigación definieron que la clasificación de las galaxias con líneas de emisión se puede definir en cinco criterios:

- Cada relación debe estar formada por líneas fuertes que sean fáciles de medir en espectros típicos.
- Se deben evitar las líneas que están mezcladas con otras líneas porque la naturaleza, un tanto subjetiva, del procedimiento de eliminación de mezcla aumenta la incertidumbre en las mediciones de flujo de estas líneas.
- La separación de longitudes de onda entre las dos líneas debe ser lo más pequeña posible para que la relación sea relativamente insensible al enrojecimiento y la calibración del flujo.
- Las proporciones que involucran una línea de un solo elemento y una línea HI Balmer deben preferirse a aquellas que involucran líneas prohibidas de diferentes elementos, porque son menos sensibles a la abundancia.

- Las líneas deben estar en una región del espectro que sea fácilmente accesible con los instrumentos, evitando las líneas ultravioleta debido a la baja sensibilidad de muchos CCD a longitudes de onda cortas.

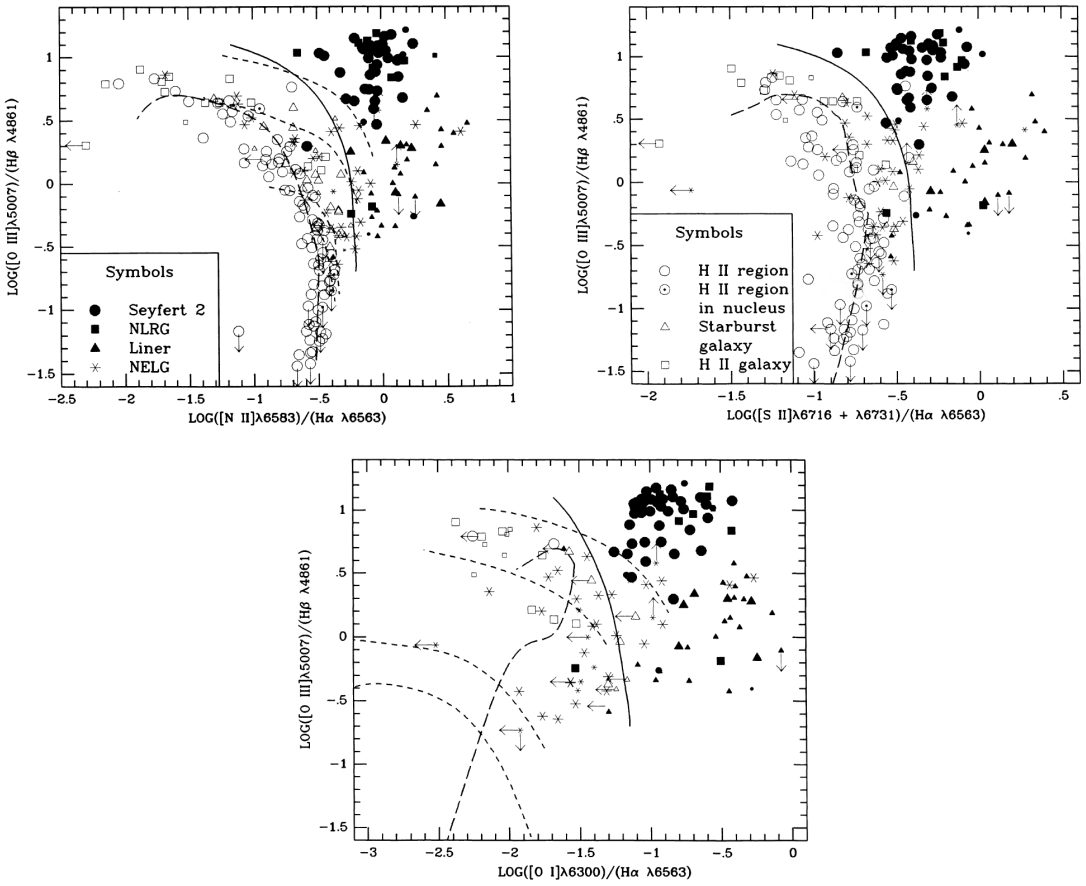


Figura 2-3: Diagramas diagnóstico para una muestra de galaxias. Las cuatro líneas de trazos cortos son modelos de región HII (Evans Dopita - 1985). La curva de trazos largos representa los modelos de región HII (McCall, Rybski Shields - 1985). La curva sólida divide los AGN de los objetos similares a regiones HII. [*Veilleux & Osterbrock*. 1987]

En 2001 Kewley y sus colaboradores parametrizaron, de manera teórica, las ecuaciones (2-1) (2-2) y (2-3), que definirían cada una de las curvas divisorias para los diferentes tipos de galaxias. Estudiaron el modelamiento teórico de los espectros observados en las galaxias con fuerte formación estelar para generar un límite superior en estos modelos a través de un diagrama de diagnóstico óptico, utilizando los límites empíricos entre ([NII]λ6584Å/H$\alpha\lambda$6563Å) vs. ([OIII]λ5007Å/H$\beta\lambda$4861Å), ([SII]λ6717Å/H$\alpha\lambda$6563Å) vs.([OIII]λ5007Å/H$\beta\lambda$4861Å) y ([OI]λ6300Å/H$\alpha\lambda$6563Å) vs. ([OIII]λ5007Å/H$\beta\lambda$4861Å). Estas líneas dividen la región de formación estelar de otros tipos de excitación. Los modelos Starburst siempre caen debajo y a la izquierda del límite calculado, como muestra la **Figura 2-4**, debido a los parámetros de ionización y de metalicidad. Por el contrario, los AGNs se ubican sobre las curvas hacia los ejes positivos [*Kewley et al.*. 2001].

$$\log\left(\frac{[OIII]\lambda5007}{H\beta}\right) = \frac{0{,}61}{log([NII]\lambda6584/H\alpha) - 0{,}47} + 1{,}19 \tag{2-1}$$

$$\log\left(\frac{[OIII]\lambda5007}{H\beta}\right) = \frac{0{,}72}{log([SII]\lambda6716/H\alpha) - 0{,}32} + 1{,}30 \tag{2-2}$$

$$\log\left(\frac{[OIII]\lambda5007}{H\beta}\right) = \frac{0{,}73}{log([OI]\lambda6300/H\alpha) + 0{,}59} + 1{,}33 \tag{2-3}$$

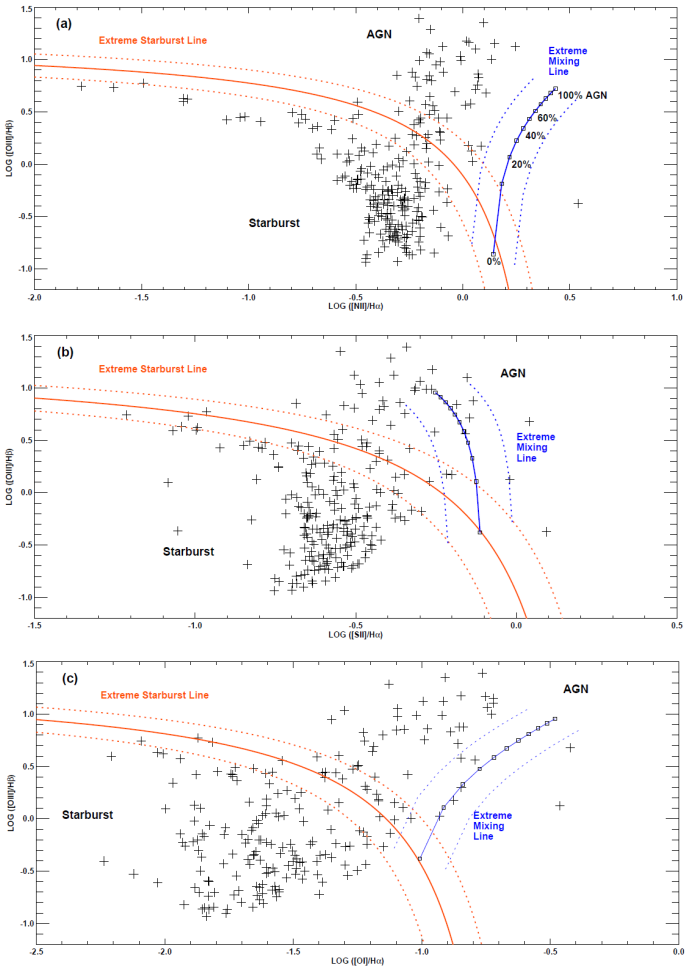


Figura 2-4: Diagramas Diagnóstico de las líneas teóricas (en rojo) para [NII], [SII], [OI]. Las líneas discontinuas representan ±0,1 e indican el rango de error del modelo. La líneas azules describen la división entre AGNs y galaxias con clasificación ‘‘ambigua’’. [*Kewley et al.*. 2001]

En el trabajo de Kauffmann del 2003 se presenta un nuevo modelo para definir la mejor curva que demarcará el límite entre los AGNs y las galaxias con formación estelar. El estudio consistió en comparar entre galaxias tipo AGN y galaxias sin núcleos activos, sus tamaños, densidades superficiales, masas y edades estelares.

La región definida entre los límites establecidos por Kewley et al. (2001) y Kauffmann et al. (2003) con la ecuación (2-4), se conoce como región compuesta y/o de transición, una zona del diagrama en el que se pueden encontrar tanto AGNs como una fracción de galaxias no-AGNs, las cuales se componen de una población estelar capaz de simular las razones de línea producidas por un núcleo activo, debido a la ionización producida por estrellas. De acuerdo con esto, en la gráfica se desplegará un grupo de galaxias tipo Starburst, otra que contiene galaxias tipo Starburst - AGN y finalmente otra donde están netamente los AGNs, como se muestra en la **Figura 2-5** [*Kauffmann et al.*. 2003].

$$\log\left(\frac{[OIII]\lambda5007}{H\beta}\right) = \frac{0{,}61}{log([NII]\lambda6584/H\alpha) - 0{,}05} + 1{,}3 \tag{2-4}$$

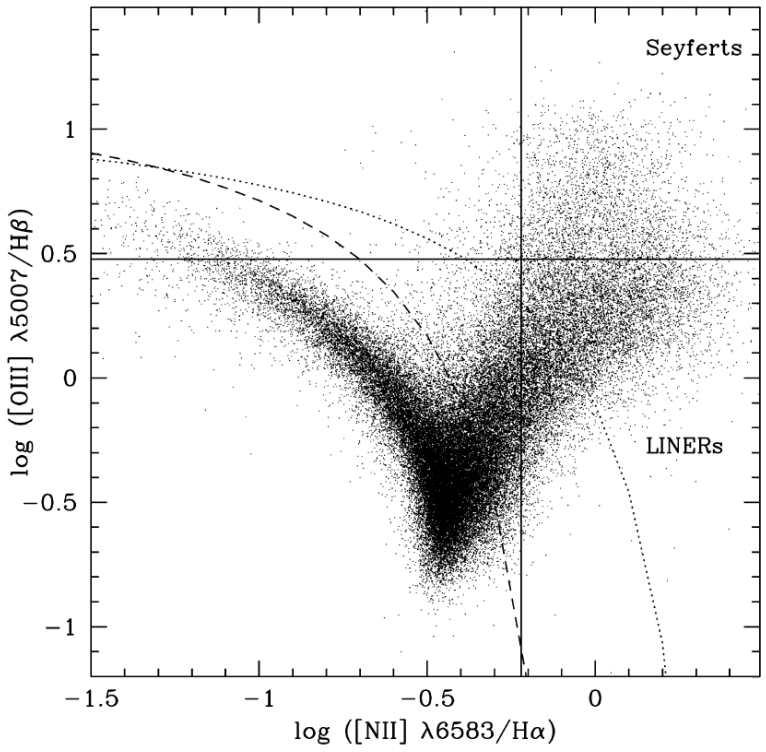


Figura 2-5: Diagrama BPT que traza la relación de flujo de la línea de emisión [OIII]/Hβ frente a la relación [NII]/Hα para 55.757 objetos. La curva punteada muestra la demarcación entre galaxias con estallido estelar y AGN, la curva discontinua muestra la demarcación revisada (2-4). Por encima de la curva discontinua se encuentran un total de 22.623 galaxias. [*Kauffmann et al.*. 2003]

Kewley en el año 2006 realizó un estudio detallado de fuentes del SDSS, proponiendo un nuevo conjunto de criterios para distinguir las galaxias Seyfert de las tipo LINER. Estos criterios se basan en un mapeo empírico de la bimodalidad observada en los diagramas [OIII]/Hβ vs. [OI]/Hα y [SII]/Hα [*Kewley et al.*. 2006], pero es en 2010 donde [*Cid Fernandes et al.*. 2010b] y su equipo derivan las ecuaciones lineales que separan estos tipos de AGNs, calculando los valores de los coeficientes de una línea recta representada en la ecuación para el plano BPT en la que se efectúa la clasificación de los dos grupos, ampliando la clasificación según el cociente de líneas espectrales (Ver Figura **2-7**) [*Cid Fernandes et al.*. 2010a].

La **Figura 2-6** es un ejemplo de un diagrama de diagnóstico (BPT) de una muestra de galaxias extraídas de la base de datos del Sloan Digital Sky Survey. El eje horizontal se define como la razón log([NII]/Hα), mientras que el eje vertical se basa en los cocientes de intensidades log([OIII]/Hβ). Este modelo representa las diferentes curvas que demarca el límite entre los AGNs y las galaxias con formación estelar, de esta manera se desplegará un grupo de galaxias tipo Starburst (SF), una segunda agrupación que contiene galaxias tipo Starburst-AGN (Comp), otra donde se encuentran los AGNs tipo Seyfert (SY) y finalmente la zona de las galaxias LINER.

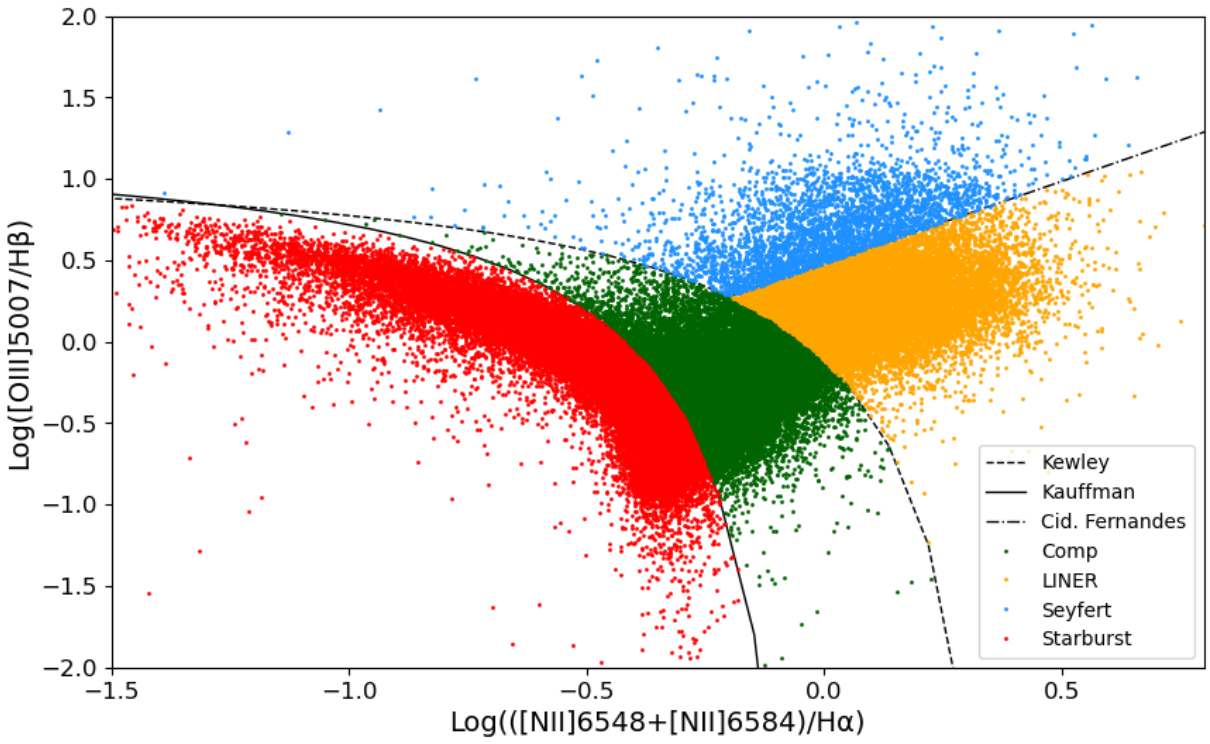


Figura 2-6: Diagrama BPT representado en Python, muestra la relación entre dos cocientes de líneas: log([OIII]λ5007/Hβ) vs. log([NII]$\lambda\lambda$6548+6584/Hα). Las galaxias en formación estelar se muestran en rojo, las compuestas en verde, Seyfert 2 en azul y las LINER en amarillo. Las curvas muestran las separaciones teóricas: la curva sólida de [*Kauffmann et al.*. 2003], la curva punteada de [*Kewley et al.*. 2001] y la línea diagonal de [*Cid Fernandes et al.*. 2010a]. Créditos: Katherine C. Soto

A continuación, se visualiza la construcción de la tabla de la **Figura 2-7** con cada una las ecuaciones que representan las curvas en el diagrama BPT, dependiendo de las líneas de emisión a analizar, esto para obtener un gráfico que permita definir los límites entre cada uno de los tipos de galaxias de la muestra.

Línea	Ecuación	Curva
NII	$\log\left(\frac{[OIII]\lambda5007}{H\beta}\right) = \frac{0{,}61}{log([NII]\lambda6584/H\alpha)-0{,}05} + 1{,}3$	SF-COMPUESTO
	$\log\left(\frac{[OIII]\lambda5007}{H\beta}\right) = \frac{0{,}61}{log([NII]\lambda6584/H\alpha)-0{,}47} + 1{,}19$	SF-LINER
	$\log\left(\frac{[OIII]\lambda5007}{H\beta}\right) = 1{,}01log([NII]\lambda6584/H\alpha) + 0{,}48$	Seyfert - LINER
SII	$\log\left(\frac{[OIII]\lambda5007}{H\beta}\right) = \frac{0{,}72}{log([SII]\lambda6716/H\alpha)-0{,}32} + 1{,}30$	SF-LINER
	$\log\left(\frac{[OIII]\lambda5007}{H\beta}\right) = 1{,}89log([SII]\lambda6716/H\alpha) + 0{,}76$	Seyfert - LINER
OI	$\log\left(\frac{[OIII]\lambda5007}{H\beta}\right) = \frac{0{,}73}{log([OI]\lambda6300/H\alpha)+0{,}59} + 1{,}33$	SF-LINER
	$\log\left(\frac{[OIII]\lambda5007}{H\beta}\right) = 1{,}18log([OI]\lambda6300/H\alpha) + 1{,}30$	Seyfert - LINER

Figura 2-7: Ecuaciones de las curvas para los diagramas BPT

2.3.2 Diagrama Diagnóstico WHAN

Generalmente, en las investigaciones se excluye a una extensa población de galaxias de líneas débiles (Weak-Line Galaxies, WLG) de los estudios estadísticos sobre galaxias de líneas de emisión (Emission-Line Galaxies, ELG), debido a la carencia de un esquema de clasificación adecuado. Esto se debe a que los diagramas de diagnóstico convencionales, como el diagrama BPT, requieren la medición de, al menos, cuatro líneas de emisión.
Este nuevo diagrama de clasificación llamado WHAN, propuesto por Cid Fernandes y su equipo, tiene como objetivo abordar esta limitación al redefinir las fronteras mediante la transposición de las líneas divisorias entre galaxias con formación estelar (SF) y núcleos galácticos activos (AGN), así como entre galaxias tipo Seyfert y LINER, realizando así una adaptación que reclasifica un número significativo de fuentes [*Cid Fernandes et al.*. 2011].

El ancho equivalente de la línea de Hα versus el cociente de líneas de [NII]/Hα (WHAN), proporciona la muestra de la gran población de galaxias de líneas débiles que no aparecen en los diagramas tradicionales, permitiendo, a través de este modelo, la diferenciación entre dos clases muy distintas que se superponen en la región de la línea de emisión nuclear de baja

ionización (LINER). Se trata de galaxias que albergan un núcleo galáctico débilmente activo (wAGN) y "galaxias retiradas"(RG), es decir, galaxias que han dejado de formar estrellas y están ionizadas por sus estrellas calientes evolucionadas de baja masa.

Dentro del diagrama WHAN, como se visualiza en la **Figura 2-8**, se identifican cinco clases de galaxias:

- Galaxias de formación estelar (SF): log[NII]/Hα < -0,4; $W_{H\alpha}$ > 3Å
- Galaxias Seyfert: log[NII]/Hα > -0,4; $W_{H\alpha}$ > 6Å
- Galaxias LINER: log[NII]/Hα > -0,4; $W_{H\alpha}$ entre 3Å y 6Å
- Galaxias Retiradas (AGN falso): $W_{H\alpha}$ < 3Å
- Galaxias pasivas (Galaxias sin líneas): $W_{H\alpha}$ < 0,5Å; log[NII]/Hα < 0,5

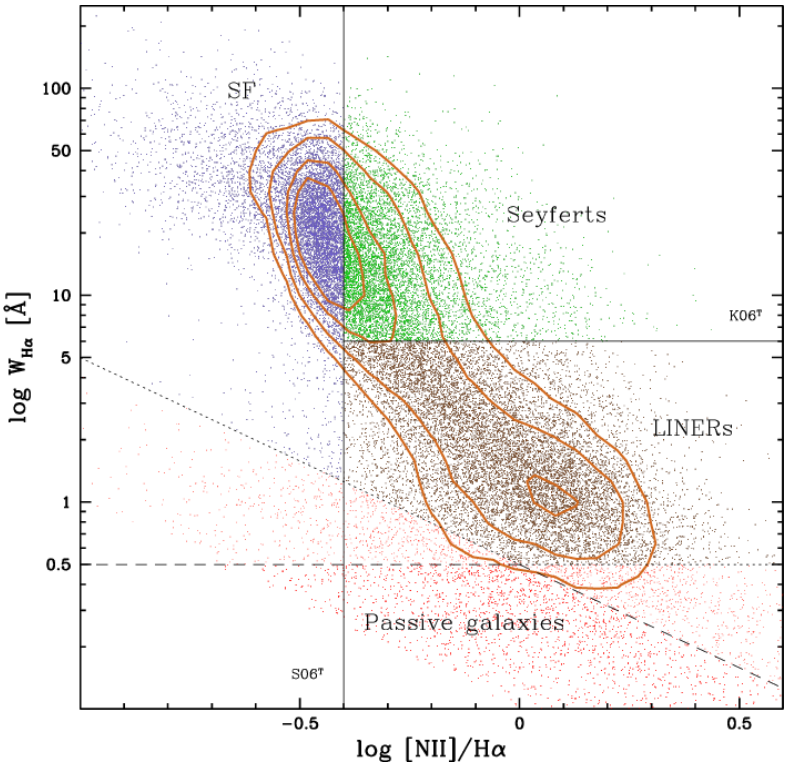


Figura 2-8: Diagrama diagnóstico WHAN. Galaxias de formación estelar en color violeta, galaxias tipo Seyfert en verde, LINER en color café y el grupo denominado galaxias pasivas en color rojo. [*Cid Fernandes et al.*. 2011]

2.3.3 Diagrama Diagnóstico Azul

Para galaxias más allá del Universo local, con corrimiento al rojo mayor que $z \approx 0{,}4$, las líneas de emisión de [NII] λ6584, [SII] $\lambda\lambda$6717+6731, Hα se desplazan al rojo fuera del rango de longitud de onda del espectro visible, por lo tanto, los diagramas diagnóstico para este tipo de objetos deben basarse en emisiones de líneas en la parte azul del espectro: [OIII] λ5007, [OII] $\lambda\lambda$3726+3729 y Hβ.

Lamareille propuso un diagrama diagnóstico que permite clasificar objetos con alto redshift, derivando un grupo de nuevas ecuaciones, como la ecuación(2-5), que define a la curva sólida de la **Figura 2-9**. Allí, las galaxias con formación estelar están bajo esta curva y las AGN arriba de la misma. En el panel derecho de la **Figura 2-9**, se evidencia un número de galaxias Seyfert 2 que caen en la región de galaxias con formación estelar, esto permite definir el límite superior a través de la ecuación (2-6), donde la región de galaxias de formación estelar se mezclan con las tipo Seyfert 2, formando una nueva región denominada SF/Sy2. El panel izquierdo de la misma figura muestra que, a diferencia de las galaxias Seyfert 2, los LINER no se mezclan significativamente con galaxias de formación estelar, para ello se define una región SF/LINER con la ecuación (2-7) [*Lamareille*. 2010].

La nueva clasificación ‘‘azul’’ mejorada de galaxias con líneas de emisión presenta un diagrama de diagnóstico donde se aplica la relación entre dos cocientes de línea: log([OIII]λ5007/Hβ) vs. log([OII]$\lambda\lambda$3726+3729/Hβ) como se representa en la **Figura 2-9**

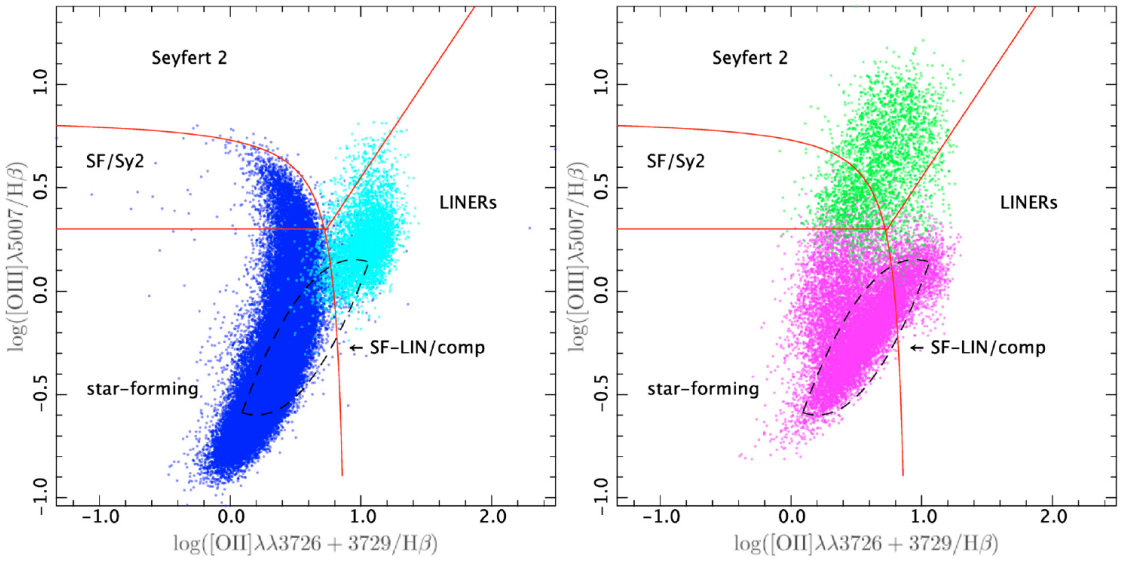


Figura 2-9: Diagrama diagnóstico azul. Permite visualizar cuatro clases de galaxias, Starburst en azul, LINER en cyan, Seyfert en verde y Compuestas en magenta. Las dos primeras clases se muestran sólo en el panel izquierdo, mientras que las dos últimas clases se muestran en el panel derecho. Las curvas rojas representan las nuevas separaciones empíricas. [*Lamareille*. 2010]

$$\log\left(\frac{[OIII]}{H\beta}\right) = \frac{0{,}11}{log([OII]/H\beta) - 0{,}92} + 0{,}85 \tag{2-5}$$

$$\log\left(\frac{[OIII]}{H\beta}\right) > 0{,}3 \tag{2-6}$$

$$\log\left(\frac{[OIII]}{H\beta}\right) = 0{,}95 * log([OII]/H\beta) - 0{,}4 \tag{2-7}$$

2.3.4 Diagrama Diagnóstico U-B

Este método de clasificación se basa en el hecho de que la mayoría de los AGNs identificados por el diagrama BPT se encuentran en galaxias rojas o con colores intermedios entre el rojo y el azul, pero pocos se encuentran en galaxias muy azules, objetos menos masivos, luminosos, con menores proporciones de bulbo-disco, lo que hace que alberguen agujeros negros más pequeños [*McLure et al.*. 2006], por lo tanto, se encontrará una fracción más baja de galaxias activas de este tipo. La formación estelar en las galaxias azules puede ocultar los AGNs, dando como resultado una clasificación errónea en el diagrama clásico.

Este modelo utiliza el índice de color U-B de las galaxias en lugar de la relación [NII]/Hα, debido a su inaccesibilidad en la ventana visible para un redshift alto (z > 0,4). El índice de color U-B se puede utilizar para rastrear la actividad del AGN, porque se correlaciona positivamente con la masa y la metalicidad, y negativamente con la tasa de formación estelar, como muestra la **Figura 2-10**. La ventaja de este método es que requiere un menor número de líneas de emisión, esto permite su aplicación en galaxias con un redshift más elevado y reduce la probabilidad de falta de completitud de líneas, sobretodo en espectros de baja razón señal-ruido [*Yan et al.*. 2011].

El eje horizontal de la **Figura 2-10** se representa con el índice de color U-B, mientras que el eje vertical se describe como log([OIII]/Hβ). Los AGNs se ubican en el panel c, parte superior derecha del diagrama, separados de las galaxias formadoras de estrellas que se encuentran en la región inferior izquierda. Las galaxias de la región de transición se superponen principalmente con las galaxias formadoras de estrellas en color U-B. La demarcación límite está dada por la ecuación (2-8), donde U-B denota el índice de color en el sistema de magnitud AB. Para valores menores a -0.1, el límite se demarca en este valor .

$$\log\left(\frac{[OIII]}{H\beta}\right) = [1{,}4 - 1{,}2(U - B), -0{,}1] \tag{2-8}$$

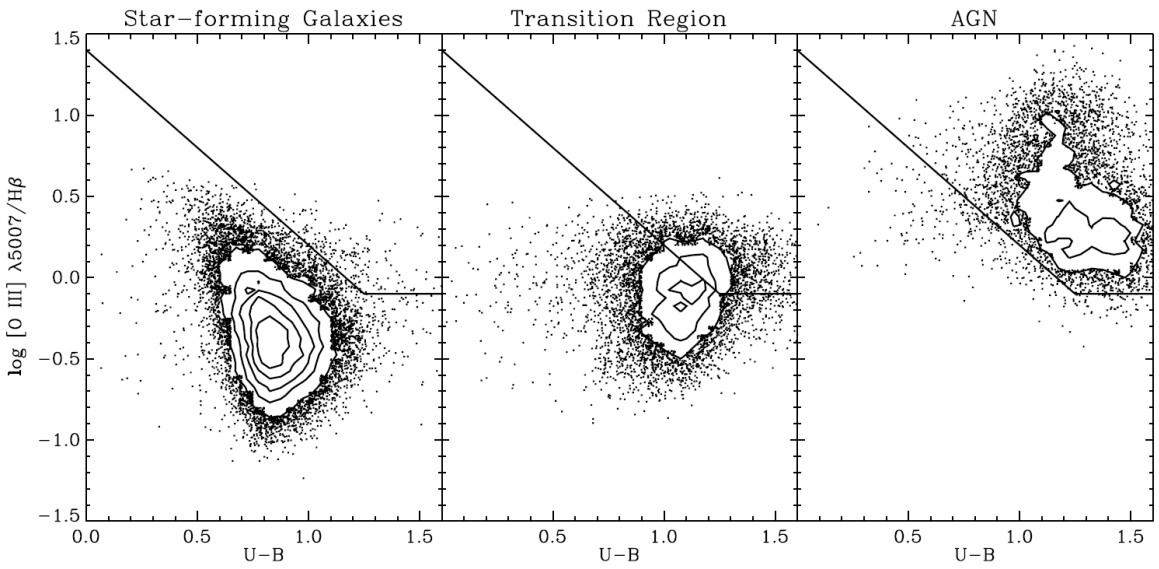


Figura 2-10: Diagrama diagnóstico U-B. El primer cuadrante muestra el grupo de galaxias de formación estelar, la segunda figura pertenece a la región de transición entre galaxias Starburst y AGN y en la tercera figura se agrupan las galaxias tipo AGN. [*Yan et al.*. 2011]

2.3.5 Diagrama Color-Color

Las investigaciones de las propiedades estadísticas de las galaxias activas han sido diseñadas para detectar tipos específicos de fuentes en función de sus características. Tal es el caso de los AGNs tipo 2 que se caracterizan por una emisión de líneas estrechas, en el que los métodos basados en diagramas diagnóstico tradicionales como el BPT permiten realizar una óptima distinción entre AGNs y actividad de formación estelar [*Baldwin et al.*. 1981]. Por otra parte, los núcleos activos de tipo 1 se identifican, generalmente, por la presencia de líneas de emisión anchas y angostas en los espectros, sin embargo, la razón por la que los cocientes de diagnóstico clásicos no pueden utilizarse en una muestra de objetos que incluyen fuentes de este tipo 1 reside en el uso de líneas de recombinación, necesarias para normalizar la intensidad de las líneas prohibidas que sondean la temperatura y la estructura de ionización del gas, a lo cual se atribuye que estos modelos funcionen únicamente con la emisión de NLR y no puede tener en cuenta el componente BLR que tienen las galaxias Seyfert 1 [*Vaona et al.*. 2012].

Se ha demostrado que los AGN de tipo 1 pueden seleccionarse, efectivamente, mediante criterios fotométricos que comparan sus colores con los de objetos no activos, por lo que el fuerte continuo en el azul y UV, producido por la fuente central permite detectar de mejor forma la actividad nuclear, además de no estar oscurecido a lo largo de la línea de visión [*Mura et al.*. 2017].

En la **Figura 2-11** se ilustra una proyección de un espacio de parámetros en el diagrama color-color u-r vs. g-z, los cuales maximizan el efecto del continuo azul de las fuentes tipo 1

sobre el continuo estelar de otras fuentes, detectando así diferentes tipos de actividad nuclear. En el espacio de parámetros tridimensional se combina líneas espectroscópicas [O III]5007/Hβ e índices de color fotométrico, donde la distribución de AGNs puebla una secuencia separada mediante criterios fotométricos, con respecto a otras fuentes de tipo 2 que se distinguen con base en sus propiedades espectrales [*Mura et al.*. 2017].

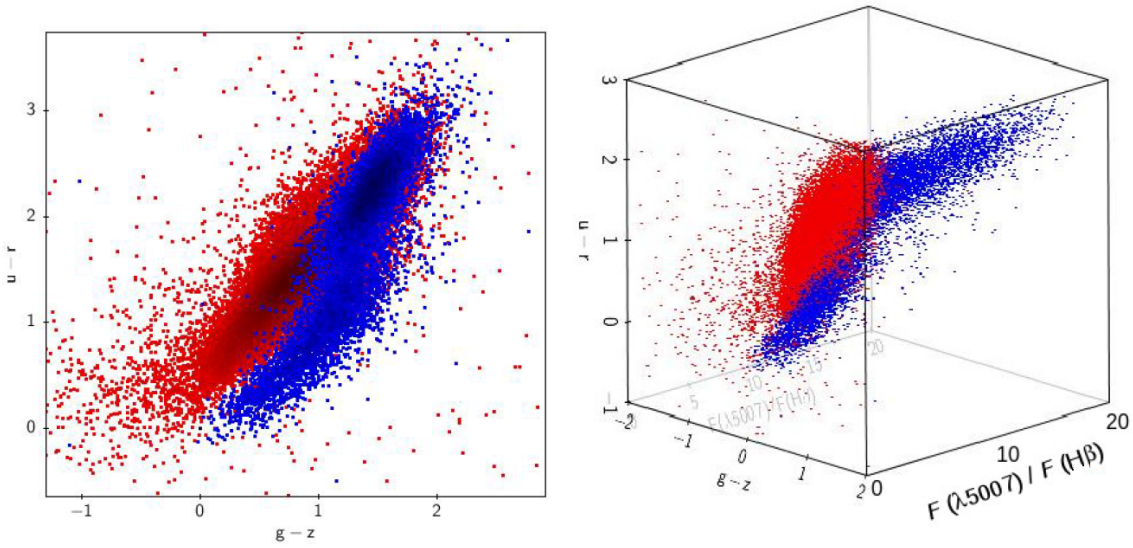


Figura 2-11: Diagrama Color-Color (panel izquierdo). Distribución de objetos tipo AGN (puntos azules) y formación de estrellas (puntos rojos), en el panel derecho se visualiza el espacio de parámetros tridimensional que combina la fotometría de dos colores con la relación de líneas espectrales [O III]/Hβ. [*Mura et al.*. 2017]

3 Machine Learning

Con el continuo progreso de las herramientas computacionales, se ha logrado ampliar la capacidad para almacenar, procesar y acceder a grandes volúmenes de datos desde ubicaciones remotas a través de redes informáticas. En la actualidad, la mayoría de los dispositivos de adquisición de datos son digitales, garantizando así la fiabilidad de la información registrada. Este avance tecnológico ha llevado al desarrollo del aprendizaje automático, una disciplina intrínseca de la inteligencia artificial.

La inteligencia artificial implica la capacidad de aprendizaje en entornos dinámicos. Un sistema verdaderamente inteligente debe tener la capacidad de adaptarse y aprender de manera continua a medida que el entorno evoluciona. La flexibilidad del aprendizaje automático radica en su capacidad para ajustarse a cambios sin requerir que los diseñadores prevean y proporcionen soluciones para todas las situaciones posibles.

El aprendizaje automático desempeña un papel crucial en la resolución de diversos problemas, para optimizar un criterio de rendimiento utilizando datos de ejemplo o experiencias previas. Se centra en el desarrollo y aplicación de algoritmos informáticos que permitan el reconocimiento de patrones y la clasificación de datos, su objetivo principal es lograr reconocer patrones de datos complejos y luego realizar aproximaciones que presenten resultados óptimos. Los modelos generados mediante aprendizaje automático pueden tener objetivos predictivos para realizar predicciones futuras, descriptivos para obtener conocimientos a partir de los datos, o incluso una combinación de ambos. La construcción de estos modelos matemáticos se apoya en la teoría estadística, ya que la tarea principal es realizar inferencias a partir de una muestra representativa [*Alpaydin.* 2014]. Algunos ejemplos de aplicaciones de aprendizaje automático son aplicaciones de aprendizaje, clasificación, regresión, aprendizaje no supervisado y aprendizaje por refuerzo.

3.1 Aprendizaje Supervisado

El aprendizaje supervisado se centra en adquirir información sobre la relación entrada - salida de un sistema mediante un conjunto específico de muestras de entrenamiento, también conocidas como datos de entrenamiento etiquetados o datos supervisados, que se caracterizan por incluir información de salida. Este enfoque busca la construcción de sistemas artificiales capaces de aprender el mapeo entre la entrada y la salida. Su objetivo principal es prever la respuesta del sistema ante nuevas características iniciales, si la salida implica un conjunto finito de valores discretos, indicando etiquetas de clase, el aprendizaje supervisado conduce a la clasificación de datos de entrada, en el caso de salidas continuas, se produce una regresión de la entrada. La representación de datos se realiza mediante parámetros del modelo de aprendizaje, cuando estos parámetros no están directamente disponibles en las muestras de entrenamiento, el sistema de aprendizaje debe llevar a cabo un proceso de estimación para obtenerlos [*Liu & Wu*. 2012].

3.1.1 Árbol de Decisión

Un árbol de decisión es un algoritmo de aprendizaje supervisado, no paramétrico, que se utiliza tanto para tareas de clasificación como de regresión. Tiene una estructura de árbol jerárquica, que consta de un nodo raíz, ramas, nodos internos y nodos hoja. Es una herramienta estructural para elegir opciones con base en criterios preestablecidos, utilizando un método de bifurcación para representar los resultados posibles al ejecutar una decisión. Dentro del árbol se generan nodos que representan variables específicas y en las ramas se puede observar el resultado de las pruebas [*Song YY*. 2015].

Modelo de un árbol de decisión

- **Nodos:** Existen tres tipos de nodos en un árbol de decisión. El nodo raíz, también llamado nodo de decisión, inicia el proceso de subdivisión de registros en subconjuntos mutuamente excluyentes. Los nodos internos, representando opciones en la estructura del árbol, están conectados al nodo padre y a nodos secundarios o nodos hoja. Los nodos hoja, finales en la jerarquía, representan el resultado final de decisiones o eventos.
- **Ramas:** Las ramas representan resultados de nodos raíz e internos. Una jerarquía de ramas forma un modelo de árbol de decisión. Cada ruta desde el nodo raíz hasta un nodo hoja representa una regla de clasificación. Dichas reglas se expresan como ‘‘si-entonces’’, proporcionando una estructura interpretable.
- **División:** Las variables de entrada, relacionadas con la variable objetivo, dividen nodos principales en subconjuntos más puros. Variables discretas o continuas se

utilizan y se seleccionan basándose en características como entropía, índice de Gini, ganancia de información, entre otros. El proceso continúa hasta cumplir con criterios de homogeneidad o de parada.

- **Parada:** Para evitar modelos sobreajustados, se aplican reglas de parada, como el número mínimo de registros en una hoja o en un nodo antes de la división. La complejidad y solidez del modelo se equilibran, seleccionando parámetros según el objetivo del análisis y características del conjunto de datos.
- **Poda:** Permite construir un árbol grande y luego eliminar nodos menos informativos para lograr el tamaño óptimo. Se selecciona el mejor subárbol posible considerando la proporción de registros con error de predicción o utilizando métodos como validación cruzada.

3.2 Aprendizaje No Supervisado

El aprendizaje no supervisado, dentro del campo del aprendizaje automático, representa un paradigma donde los algoritmos son entrenados en conjuntos de datos que carecen de etiquetas. Esta ausencia de supervisión implica que no hay información previa sobre las salidas deseadas o clasificaciones para guiar el proceso de entrenamiento. El objetivo principal radica en descubrir patrones, estructuras o relaciones inherentes en los datos mismos, sin depender de la orientación proporcionada por salidas etiquetadas [*Naeem et al.*. 2023].

Al prescindir de la necesidad de datos etiquetados, el aprendizaje no supervisado puede ser considerado en diversas aplicaciones, desde la exploración de patrones en grandes conjuntos de datos hasta la identificación de estructuras complejas dentro de la información. Este enfoque versátil ofrece una perspectiva valiosa para la comprensión de datos complejos y representa un componente esencial en el continuo avance del aprendizaje automático.

3.2.1 Clustering (K-Means)

El modelo K-means, una herramienta fundamental en el ámbito de reconocimiento de patrones, se destaca por su capacidad para descubrir grupos distintos en conjuntos de datos no etiquetados. Este enfoque no supervisado asigna cada punto a un grupo específico, buscando minimizar el índice de rendimiento del clúster, el error cuadrático y el criterio de error. El algoritmo trabaja iterativamente para asignar a cada punto uno de los "K" grupos basado en sus características [*Zubair, M.*. 2022].

Uno de los principales objetivos radica en la determinación de los centroides iniciales de cada

uno de los grupos. La técnica de distancia mínima al cuadrado distribuye cada punto a los grupos o subgrupos más cercanos, y su tarea es la de establecer la posición óptima de los centroides desde la primera iteración, en la cual se eligen algunos puntos como representantes iniciales de los grupos, generalmente los primeros K puntos de muestra. Luego se agrupan los puntos restantes según el criterio de distancia mínima, generando una clasificación inicial. Si la clasificación no es razonable, se ajusta mediante la recalculación de cada punto focal del grupo, repitiendo este proceso hasta obtener una clasificación satisfactoria [*Li & Wu*. 2012].

La **Figura 3-1** es la representación de este algoritmo que sigue las siguientes fases:

- **Inicialización:** Se elige la localización de los centroides de los K grupos aleatoriamente.
- **Asignación:** Se asigna cada dato al centroide más cercano.
- **Actualización:** Se actualiza la posición del centroide al mínimo de la suma de las distancias cuadradas desde la media aritmética de las posiciones de los datos asignados al grupo.

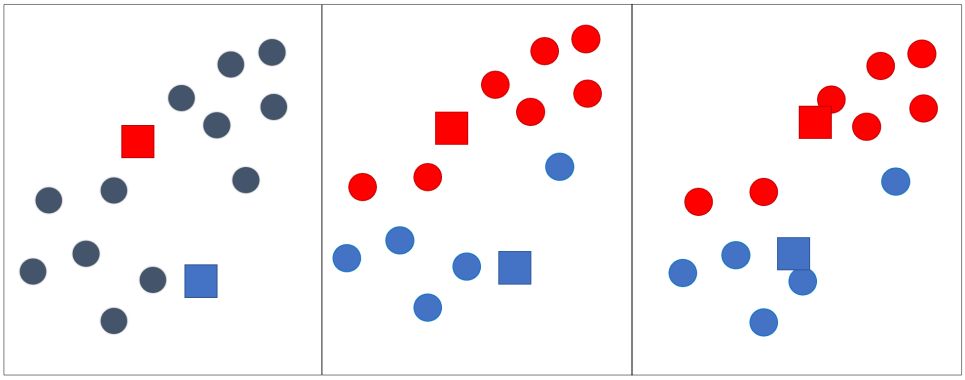


Figura 3-1: Algoritmo de Clustering: Representación de distribución de los centroides de un grupo aleatorio de datos a través de las tres fases.

3.3 Deep Learning

El Aprendizaje Profundo es una rama destacada del campo del aprendizaje automático, se fundamenta en un conjunto especializado de algoritmos diseñados para modelar abstracciones de alto nivel de datos complejos a través de redes neuronales artificiales (RNA). Su evolución se remonta a la propuesta inicial de una RNA por [*McCulloch & Pitts*. 1943], quienes idearon un modelo computacional para simular la colaboración de neuronas biológicas en la realización de cálculos complejos mediante lógica proposicional.

El Deep Learning se distingue por la presencia de múltiples capas de procesamiento, compuestas por funciones lineales y no lineales, que se adentran en la identificación de patrones

en estas capas, permitiendo que las computadoras desarrollen comprensión a partir de la información sin depender de estructuras predefinidas. Esta disciplina integral se centra en la investigación de redes neuronales, inteligencia artificial, modelado gráfico, identificación, optimización de patrones, y procesamiento de señales. Las RNA más destacadas en la actualidad incluyen el Perceptrón Multicapa, las Redes Neuronales Convolucionales, las Redes Neuronales Recurrentes y las Redes Generativas Antagónicas [*Chagas*. 2019]. Desde la primera RNA hasta la actualidad, se ha observado un crecimiento exponencial en la diversidad de tipos y arquitecturas de RNA de Aprendizaje Profundo.

Las investigaciones actuales en este tipo de aprendizaje se centran en mejorar las técnicas y aplicaciones. El acceso a nuevas capacidades de cómputo, como las GPUs, ha potenciado significativamente el rendimiento, permitiendo desarrollos a través de de lenguajes de programación como Python.

3.3.1 Redes Neuronales

Las Redes Neuronales Artificiales (RNA) se definen como modelos matemáticos que encuentran su inspiración en el complejo comportamiento biológico de las neuronas y la estructura cerebral. Estos modelos han demostrado su utilidad al abordar una amplia gama de problemas, aprovechando su notable flexibilidad para adaptarse a diversas tareas. Similar a la arquitectura de un sistema neuronal biológico, las RNA se fundamentan en elementos clave llamados neuronas, cada neurona artificial es un dispositivo de cálculo simple que genera una respuesta única a partir de un conjunto de datos de entrada. Estas neuronas se organizan en la red en niveles o capas, desempeñando funciones específicas según su posición [*Centeno Franco*. 2019].

La primera capa, conocida como capa de entrada, desempeña el papel de recibir directamente la información del entorno exterior, incorporándola al sistema neuronal. En contraste, las capas ocultas, situadas internamente en la red, son responsables del procesamiento de los datos de entrada, realizando cálculos complejos para extraer patrones y características relevantes. Por último, la capa de salida transfiere la información procesada hacia el exterior, ofreciendo la respuesta final de la red.

La configuración de una RNA puede variar, con la posibilidad de tener varias capas ocultas o incluso ninguna. Los enlaces sinápticos, representados por las flechas que conectan las neuronas, indican el flujo de la señal a través de la red y llevan asociado un peso sináptico específico. Cuando la salida de una neurona se dirige hacia múltiples neuronas en la siguiente capa, cada una de estas recibe la salida neta de la neurona anterior. El número total de capas en una RNA se determina sumando las capas ocultas más la capa de salida [*Bishop*. 2013]. Este diseño jerárquico y estratificado permite a las RNA abordar problemas complejos mediante la representación y procesamiento de información de manera eficiente.

3.3.2 Estructura de una Red Neuronal

Las redes neuronales son modelos creados para ordenar operaciones matemáticas siguiendo una determinada estructura representada mediante el uso de capas (layers), formadas, a su vez, por neuronas que realizan operaciones y están conectadas a la capa anterior y la capa siguiente mediante pesos.

Perceptrón

El perceptrón es la unidad básica de procesamiento, como modelo de neurona artificial, toma múltiples entradas que pueden provenir del exterior o pueden ser salidas de otras neuronas. Asociado a cada entrada, $x_i \in R$, i = 1,...,n, $w_i \in R$, es el peso de conexión o peso sináptico, sumando a ellos un bias o sesgo b. La salida, y, en el caso más simple es una suma ponderada de las entradas [*Alpaydin*. 2014].
La ecuación (3-1) es la representación matemática de la función del perceptrón, mientras que la **Figura 3-2** es la representación visual.

$$y = \sum_{i=1}^{n} = x_i w_i + b \tag{3-1}$$

El perceptrón está conformado por una serie de componentes:

- **Entradas:** Las entradas en el algoritmo se entienden como $x_1, x_2, x_3, ..., x_i$ y así sucesivamente, las cuales representan los valores de las características que ingresan a la neurona.
- **Pesos:** Son parámetros ajustables que se utilizan para ponderar las entradas de cada neurona, cada conexión entre dos neuronas tiene un peso asociado que determina la fuerza de la conexión y su impacto en la salida de la neurona receptora. Básicamente representan la influencia relativa de la entrada por la cual se multiplica x_i. Los pesos ofrecen un valor aleatorio preliminar en el inicio del aprendizaje del algoritmo, los valores de los pesos se actualizan en cada iteración para disminuir el error. Estos se representan principalmente como $w_1, w_2, w_3, ..., w_i$.
- **Bias:** Su función principal es desplazar el resultado obtenido a través de la función de activación, lo que permite una mayor flexibilidad en la representación de los patrones de entrada. Se le llama sesgo, ya que controla qué tan predispuesta está la neurona a disparar un 1 o un 0 independientemente de los pesos. Un sesgo alto hace que la neurona requiera una entrada más alta para generar una salida de 1 y viceversa.
- **Suma ponderada:** Es la proliferación de cada valor de entrada o característica asociada con el valor de paso correspondiente.

- **Función de activación:** Es una función matemática que determina la salida de una neurona o de un conjunto de neuronas en función de sus entradas ponderadas. Esta función introduce la no linealidad en el modelo, lo que permite a la red aprender patrones complejos y realizar tareas no lineales.

- **Salida:** La suma ponderada se pasa a la función de activación y el valor obtenido después del proceso será el resultado de la salida predicha. Si la función de activación está por debajo de un umbral determinado, ninguna salida se pasa a la neurona subsiguiente.

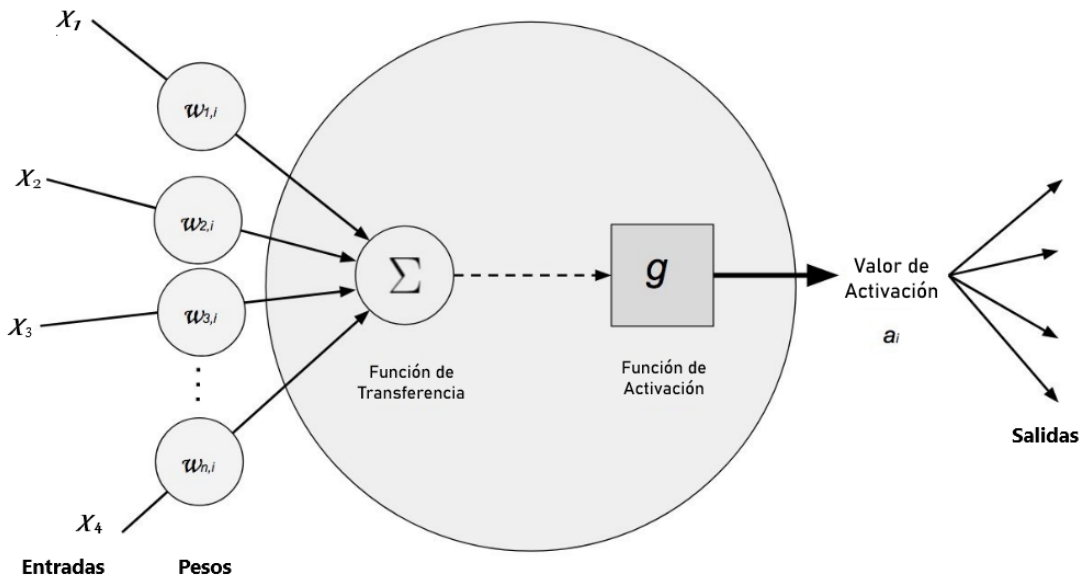


Figura 3-2: Esquema del modelo de un perceptrón con 4 entradas, pesos, la función de transferencia, la función de activación y sus respectivas salidas. (Adaptación). Fuente: Deep Learning A Practitioner's Approach, Josh Patterson and Adam Gibson.

3.3.3 Funciones de Activación

Las funciones de activación son una función matemática que se aplica a la salida de una neurona y se utilizan para introducir no linealidad en la red neuronal, lo que le permite a la red aprender y representar relaciones no lineales en los datos. Existe un gran número de funciones que se han ido proponiendo y perfeccionando a través del tiempo. Como se muestra en la **Figura 3-3**, estas son las funciones de activación generalmente utilizadas dependiendo del tipo de problema que se quiera resolver. Las funciones más comunes se describen a continuación:

Figura 3-3: Funciones de activación de redes neuronales. Tomado de https://www.v7labs.com/

Función Lineal

La función lineal o función identidad, es una función matemática simple que asigna cada valor de entrada directamente a su valor de salida sin aplicar ninguna transformación, es proporcional a la entrada. La forma más básica de una función lineal se define en la ecuación (3-2). Gráficamente, esta función representa una línea recta que pasa por el origen con una pendiente de 45 grados. En el contexto de las redes neuronales, esta función se utiliza como función de activación en capas de salida cuando se quiere realizar una regresión, ya que la salida debe ser una combinación lineal de las entradas, sin embargo, en capas ocultas, se utilizan funciones no lineales, para permitir que la red aprenda patrones más complejos.

$$f(x) = x \tag{3-2}$$

Función Sigmoide

También conocida como función logística, es una función matemática que toma cualquier número real como entrada y produce una salida en el rango de 0 a 1. Una función sigmoidea es una función acotada y diferenciable que no es decreciente y tiene exactamente un punto de inflexión. Esta suaviza las transiciones entre las dos regiones extremas, lo que ayuda en la convergencia durante el entrenamiento de modelos [*Lederer*. 2021]. Generalmente se utiliza en problemas de clasificación binaria, donde la salida se interpreta como la probabilidad de pertenencia a una clase. Su ecuación está representada en (3-3).

$$\sigma(x) = \frac{1}{1 + e^x} \tag{3-3}$$

Función Tangente Hiperbólica

La función Tanh produce valores en el rango de -1 a 1, lo que la hace útil en problemas donde se requiere una salida que pueda variar tanto positiva como negativamente. Es una función simétrica que comparte algunas de las propiedades beneficiosas de la sigmoide, como la suavidad y la derivada simple [*Lederer*. 2021]. Se describe a través de la ecuación (3-4).

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{3-4}$$

Función ReLU

Llamada Unidad Lineal Rectificada, por sus siglas en íngles Rectified Linear Unit, la función ReLU retorna el valor x si x es mayor que cero, es decir permite el paso de todos los valores positivos sin cambiarlos y activa la neurona, mientras que para valores negativos la función retorna 0. La función ReLU introduce no linealidades en el modelo, lo que permite a la red aprender patrones más complejos y no lineales en los datos [*Lederer*. 2021]. Generalmente es utilizada en problemas de clasificación multiclase, para la cual, la ecuación (3-5) representa esta función.

$$f(x) = \begin{Bmatrix} 0 & si & x < 0 \\ x & si & x \geq 0 \end{Bmatrix} \tag{3-5}$$

Las ventajas de utilizar ReLU como función de activación son las siguientes:

- Dado que solo se activa una cierta cantidad de neuronas, la función ReLU es mucho más eficiente computacionalmente en comparación con otras funciones.

- ReLU acelera la convergencia del descenso del gradiente hacia el mínimo global de la función de pérdida debido a su propiedad lineal y no saturante.
- La función activa solo las neuronas con entradas positivas, lo que puede conducir a una representación más eficiente de los datos.

Una de limitaciones que enfrenta esta función es el llamado problema de "Dying ReLU" o "Muerte de ReLU", esto se presenta debido a que todos aquellos valores que sean negativos siempre tomarán un valor de cero. El lado negativo del gráfico hace que el valor del gradiente sea cero, por lo tanto, durante el proceso de retropropagación, los pesos y sesgos de algunas neuronas no se actualizan, esto puede crear neuronas muertas que nunca se activan durante el entrenamiento.

Función Leaky ReLU

Esta función de activación es una variante de la función ReLU, la cual utiliza un valor pequeño y no nulo para las entradas negativas. En esta definición, x es la entrada a la función y α es un parámetro variable que determina la pendiente de la función para valores negativos, generalmente el valor de $\alpha = 0{,}01$ o se puede seleccionar durante el entrenamiento como un parámetro aprendido como se determina en la ecuación (3-6).

$$f(x) = \begin{Bmatrix} \alpha x & si & x < 0 \\ x & si & x \geq 0 \end{Bmatrix} \tag{3-6}$$

Las ventajas de Leaky ReLU son las mismas que las de ReLU, además del hecho de que permite la propagación hacia atrás, incluso para valores de entrada negativos. Al realizar esta modificación para valores negativos, el gradiente del lado izquierdo del gráfico resulta ser un valor distinto de cero, por lo tanto, ya no se encontrarán neuronas muertas en esa región [*Lederer*. 2021]. Una desventaja de este tipo de función es que es posible que las predicciones no sean consistentes para valores de entrada negativos, además de que el gradiente de estos valores es un número pequeño que hace que el aprendizaje de los parámetros del modelo lleve mayor tiempo de procesamiento.

Función Softmax

La función Softmax es una función de activación utilizada comúnmente en la capa de salida de una red neuronal cuando se trata de problemas de clasificación con múltiples clases. Esta función toma un vector de números reales y los transforma en un vector de probabilidades, donde cada elemento del vector de salida representa la probabilidad de pertenencia a una clase específica, de tal manera que la sumatoria de todas las probabilidades de la salida es

igual a 1 [*Nwankpa et al.*. 2020]. La función amplifica las diferencias entre las probabilidades, lo que significa que la clase con la probabilidad más alta se vuelve más evidente y fácil de distinguir.

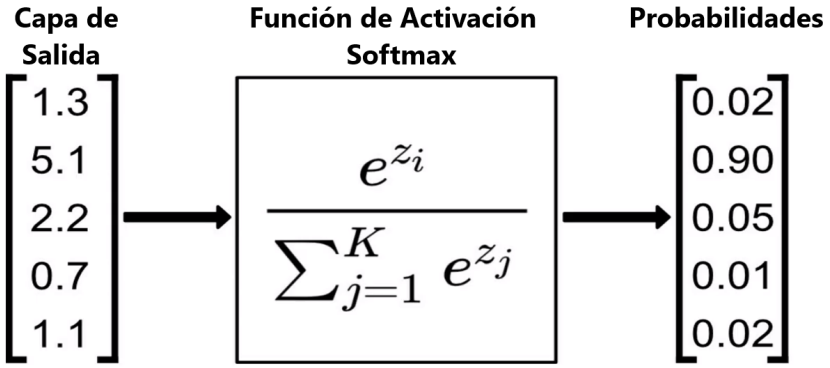


Figura 3-4: Representación en bloques de la función de activación Softmax.

La fórmula matemática de la función Softmax para un vector z de K elementos se muestra en la **Figura 3-4**, cuya representación evidencia los valores de entrada que corresponden a la capa de salida de la neurona anterior, seguido del bloque con la función matemática de Softmax, dando como resultado el tercer bloque de datos con cada una de las probabilidades. En la ecuación z_i es el i-ésimo elemento del vector z, y la función Softmax calcula la probabilidad de que el elemento i-ésimo sea la clase correcta.

3.4 Librerías de Machine Learning

3.4.1 Scikit-Learn

Scikit-Learn es una biblioteca de aprendizaje automático de código abierto para el lenguaje de programación Python de última generación, creado para problemas supervisados y no supervisados. Proporciona una amplia variedad de herramientas y algoritmos para tareas comunes de aprendizaje automático, como clasificación, regresión, agrupación, reducción de dimensionalidad y selección de modelos [*Pedregosa et al.*. 2018].

Algunas características principales que lo definen son:

- Scikit-Learn proporciona una interfaz fácil y coherente de utilizar para varios algoritmos de aprendizaje automático, algoritmos tales como máquinas de soporte vectorial (SVM), K-means, regresión lineal y logística, árboles de decisión, entre otros.

- Ofrece herramientas de preprocesamiento de datos, como la transformación y normalización de características, manejo de valores faltantes, codificación de variables categóricas.
- Proporciona métricas y herramientas para evaluar el rendimiento de los modelos, así como técnicas de validación cruzada.
- Incluye utilidades para realizar la selección de modelos y optimización de hiperparámetros, ayudando a encontrar la mejor configuración para un modelo específico.

3.4.2 Keras

Keras es una biblioteca de redes neuronales artificiales de código abierto, desarrollada en Python. Se ha diseñado para ser modular, extensible y fácil de usar, capaz de construir, a través de bloques, la arquitectura de cada red neuronal, incluyendo redes convolucionales y recurrentes, que son las que permiten entrenar modelos deep learning. Keras proporciona una interfaz de alto nivel para definir modelos de RNA, los cuales utilizan diferentes motores de backend, siendo TensorFlow y Theano los más comunes [*John Joseph et al.*. 2021].

Características clave de Keras:

- Keras aporta una interfaz que facilita la definición de modelos de aprendizaje profundo. Permite construir modelos de manera rápida y sencilla utilizando una sintaxis simple y modular.
- Aunque TensorFlow es el backend predeterminado, Keras también es compatible con otros motores de backend, como Theano y Microsoft Cognitive Toolkit (CNTK).
- Proporciona una variedad de capas predefinidas, como capas densas, capas de convolución, capas de recurrencia, capas personalizadas, que permiten la construcción de modelos secuenciales y modelos funcionales.
- Se simplifica el proceso de entrenamiento y evaluación de modelos, ofreciendo funciones integradas para la compilación de modelos, la configuración de funciones de pérdida y optimización, y la evaluación del rendimiento del modelo.

3.5 Hiperparámetros

Los hiperparámetros son variables de configuración utilizados para administrar el entrenamiento de modelos de Machine Learning. Cada grupo de datos y cada modelo necesitan

un conjunto diferente de hiperparámetros, los cuales se determinan mediante la realización de múltiples pruebas que se ejecutan a través del modelo para encontrar cuáles de ellos representan un proceso optimizado. Entre los ejemplos de hiperparámetros se incluyen el número de nodos y capas de una red neuronal y el número de ramificaciones de un árbol de decisiones, la arquitectura del modelo, la tasa de aprendizaje y los inicializadores, entre otros [*Hutter et al.*. 2019].

A continuación, se especifican los hiperparámetros elegidos para el modelo de red neuronal que se está trabajando en esta investigación.

Las estructuras de datos centrales de Keras se constituyen de capas y modelos que constituyen el algoritmo de red neuronal. El tipo de modelo más simple utilizado es el llamado secuencial, ya que se tiene una red de neuronas consecutivas.
Para iniciar, definir y compilar el modelo, la clase secuencial permite construir la red neuronal apilando las diferentes capas usando el método **"add"** y la clase **"Dense"**, permitiendo que cada una de las capas de la red neuronal se interconecten unas a otras.

En la primera capa densa en un modelo se pasan cuatro parámetros, iniciando con el número de neuronas que se requieren para esa capa, segundo, la dimensión de entrada debe especificarse estableciendo el parámetro "input_dim" que es del mismo tamaño de la cantidad de entradas, tercero se selecciona la función de activación y por último el inicializador de pesos [*John Joseph et al.*. 2021].

3.5.1 Kernel Inicializador

Los inicializadores de peso representan cómo se establecen los valores iniciales de la matriz de peso de una capa de red neuronal. Los pesos son parámetros ajustables que se utilizan durante el proceso de entrenamiento para aprender patrones y representaciones útiles de los datos. Cuando el algoritmo de aprendizaje profundo se propuso, era común iniciar ponderaciones con ruido gaussiano, estableciendo la media igual a cero y la desviación estándar a 0.01, pero esta forma de inicialización de pesos no era óptima para aprendizaje profundo debido a problemas, como desaparición del gradiente o neuronas muertas [*Li et al.*. 2020].

La elección de cómo inicializar los pesos de una red neuronal puede afectar significativamente el rendimiento del modelo y su capacidad para converger durante el entrenamiento. Diferentes kernel inicializadores pueden influir en la velocidad de convergencia y en la calidad de las soluciones encontradas. En general, estos métodos se pueden dividir en dos categorías: inicialización de ceros y unos e inicialización aleatoria.

1. **Inicialización de ceros y unos:** Con todos los pesos inicializados en 0 o 1, todos son iguales y la activación en las neuronas también es la misma, de esa manera, la derivada de la función de pérdida es la misma para cada peso en una matriz de pesos

de una capa. En todas las iteraciones, las capas ocultas se vuelven simétricas, cada neurona de la capa calcula la misma función, por lo que el modelo se comporta como un modelo lineal.

2. **Inicialización aleatoria:** Todos los valores de la matriz de peso se establecen en números aleatorios, generalmente de una distribución normal o uniforme. El problema con este tipo de inicialización son los gradientes que desvanecen, donde la actualización de peso es menor, lo que resulta en una convergencia más lenta, mientras que en gradientes explosivos, los gradientes grandes pueden resultar en una oscilación alrededor del valor óptimo.

Para redes profundas, se pueden utilizar heurísticas para inicializar los pesos según la función de activación no lineal. La heurística establece la varianza de la distribución normal en k/n, donde k es un valor constante que depende de la función de activación y n es el número de nodos de entrada. A continuación, se enumeran diferentes inicializadores de peso donde fan_{in} es el número de unidades de entrada en el tensor de peso y fan_{out} es el número de unidades de salida en el tensor de peso.

- **Inicializador Xavier/Glorot Normal:** Propuesta en 2010 por Glorot y Bengio, esta inicialización ajusta los pesos de acuerdo con la cantidad de neuronas de entrada y salida en una capa. Se basa en la idea de mantener la varianza constante a lo largo de las capas para evitar el desvanecimiento o la explosión del gradiente [*Glorot & Bengio*. 2010]. Se extrae cada peso W de una distribución normal con una media de 0 y una desviación estándar igual a 2, dividida por el número de entradas más el número de salidas para la transformación. La ecuación (3-7) representa el inicializador $Xavier\ Normal$

$$W = N(\mu, \sigma)$$

$$\mu = 0 \qquad \sigma = \sqrt{\frac{2}{fan_{in} + fan_{out}}} \tag{3-7}$$

- **Inicializador He Normal:** En 2015, [*He et al.*. 2015] evidenció que el inicializador Glorot no funciona de la mejor manera con la función de activación de ReLU y se propuso la fórmula aumentando la escala en $\sqrt{2}$. Se define distribuyendo los pesos iniciales de una capa de manera aleatoria según una distribución de probabilidad normal (gaussiana) con una media de cero y una varianza que depende del número de entradas a la capa. La ecuación (3-8) representa el inicializador $He\ Normal$

$$W = N(\mu, \sigma)$$

$$\mu = 0 \qquad \sigma = \sqrt{\frac{2}{fan_{in}}} \tag{3-8}$$

- **Inicialización de LeCun Normal:** Propuesta por [*Lecun et al.*. 2000], este inicializador ajusta los pesos utilizando la cantidad de neuronas de entrada, una distribución específica para generar los valores iniciales de los pesos. La ecuación (3-9) representa el inicializador *LeCun Normal*

$$W = N(\mu, \sigma)$$

$$\mu = 0 \qquad \sigma = \sqrt{\frac{1}{fan_{in}}} \tag{3-9}$$

En el modelo de compilación se especifica el proceso de aprendizaje, tomando los diferentes hiperparámetros que son definidos dependiendo el tipo de algoritmo a desarrollar.

3.5.2 Función de Pérdida

La función de pérdida de entropía cruzada o "Categorical Crossentropy" es una de las funciones de pérdida más utilizadas para entrenar modelos de redes neuronales profundas, especialmente en problemas de clasificación multiclase. Cuando se aplica a datos categóricos, esta función de pérdida corresponde a una probabilidad logarítmica, diseñada para cuantificar la diferencia entre la distribución de probabilidad de las predicciones del modelo y la distribución de probabilidad real de las etiquetas, lo que da como resultado propiedades de estimación favorables [*Gordon-Rodriguez et al.*. 2020].

La entropía cruzada categórica penaliza fuertemente las predicciones incorrectas y tiende a cero cuando la predicción se acerca al dato correcto. Es una función de pérdida comúnmente utilizada en combinación con funciones de activación como Softmax en la capa de salida de modelos de clasificación múltiple.

La fórmula de la función de entropía cruzada categórica para un problema de clasificación con N clases está representada en (3-10).

$$L(y, \hat{y_i}) = -\sum_{i=1}^{N} y_i \; log(\hat{y_i}) \tag{3-10}$$

Donde:

- y es el vector de etiquetas verdaderas (one-hot encoded), es decir, un vector binario con un 1 en la posición correspondiente a la clase real y 0 en las demás.
- $\hat{y_i}$ es el vector de probabilidades predichas por el modelo para cada clase.
- El término $y_i \; log(\hat{y_i})$ mide la discrepancia entre la predicción del modelo ($\hat{y_i}$) y el dato verdadero (y_i) para cada clase.
- La pérdida total es la suma ponderada de estas discrepancias.

3.5.3 Optimizadores

En Keras, el optimizador es un componente encargado de ajustar los pesos del modelo con base en la función de pérdida calculada durante la fase de entrenamiento. Su funcionamiento esencial se basa en calcular el gradiente de la función de coste (derivada parcial) por cada peso de la red. Para minimizar el error se modifica cada peso en la dirección negativa del gradiente, ya que se busca agilizar la convergencia de la función de coste hacia su mínimo, se multiplica el vector de gradiente por un factor llamado factor de entrenamiento.

El conjunto de métodos iterativos de reducción de la función de error, en la búsqueda de un mínimo local, son conocidos como los métodos de optimización basados en el gradiente descendente. Diferentes optimizadores utilizan estrategias y algoritmos variados para realizar esta optimización, entre los principales utilizados en las investigaciones se encuentran: Adadelta, Adagrad, Adam, Adamax, Ftrl, Nadam, RMSprop, SGD (Stochastic Gradient Descent).

ADAM

Adam (Adaptive Moment Estimation) o Estimación Adaptativa de Momentos, es un método para la optimización estocástica eficiente, que solo requiere gradientes de primer orden con poca memoria. Propuesto por [*Kingma & Ba*. 2017], Adam combina las ideas de otros dos optimizadores, el método AdaGrad y el método de RMSprop para proporcionar una eficiente optimización de parámetros. El método calcula tasas de aprendizaje adaptativo individuales para diferentes parámetros a partir de estimaciones del primer y segundo momento de los gradientes.

- **Adagrad (Adaptive Gradient Algorithm):** Un método que introduce el aprendizaje adaptativo, en que el nivel de variación de los parámetros depende de estos. Si en el descenso de gradiente se avanza en la dirección de la mayor pendiente en cada momento, se intenta que la dirección de descenso, aunque no sea la máxima, sí apunte en la dirección del mínimo [*Duchi et al.*. 2011].
- **RMSprop (Root Mean Square prop):** Utiliza una media móvil de los cuadrados del gradiente y normaliza ese valor, empleando para ello las magnitudes recientes de los gradientes anteriores [*Hinton et al.*. 2012].

Básicamente, Adam es un algoritmo de optimización que verifica la evolución de la tasa de aprendizaje o *Learning Rate* en cada iteración para encontrar el punto mínimo del error del entrenamiento. En comparación con los demás optimizadores, este método ha presentado los mejores resultados dentro de las investigaciones con clasificación múltiple. La ecuación

(3-11) representa matemáticamente este optimizador:

$$W_t = W_{t-1} - \alpha \frac{\hat{m}_t}{\sqrt{\hat{\nu}_t} + \varepsilon} \tag{3-11}$$

$\hat{m}_t = \beta_1 * \hat{m}_t + (1 - \beta_1) * \Delta W$
$\hat{\nu}_t = \beta_2 * \hat{\nu}_t + (1 - \beta_2) * \Delta W^2$

Donde:

- W_t**:** Valor de los pesos
- α**:** Tasa de aprendizaje, generalmente tiene un valor de 0.01
- $\hat{m}_t$**:** Es el término de momento adaptativo
- $\hat{\nu}_t$**:** Es el término de RMSprop adaptativo
- ε**:** Es un pequeño valor para evitar la división por cero
- m**:** Media de los pesos a lo largo del tiempo.
- ν**:** Varianza de los pesos a lo largo del tiempo.
- β_1**:** 0.9
- β_2**:** 0.999

3.5.4 Métricas

En el contexto del entrenamiento de modelos de aprendizaje automático y profundo, las métricas son medidas utilizadas para evaluar el rendimiento del modelo en tareas específicas, proporcionando información cuantitativa sobre la calidad de las predicciones del modelo. Son particularmente importantes para problemas de clasificación, donde el objetivo es asignar datos de entrada a clases o etiquetas predefinidas. Las métricas de precisión incluyen exactitud de clasificación, precisión, recuperación y puntuación F1, estas ayudan a evaluar la capacidad del modelo para clasificar correctamente los datos e identificar posibles sesgos y limitaciones [*Jierula et al.*. 2021].

Algunas de las métricas más utilizadas en problemas de clasificación incluyen:

- **Exactitud (Accuracy):** Calcula el número de predicciones correctas respecto al total de instancias.

- **Matriz de Confusión:** Muestra el número de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos.
- **Recuperación (Recall):** Mide la proporción de instancias positivas correctamente clasificadas entre todas las instancias que realmente son positivas.
- **Área Bajo la Curva ROC (AUC-ROC):** Evalúa la capacidad del modelo para discriminar entre clases positivas y negativas, variando el umbral de decisión.
- **F1-Score:** Es la media armónica de precisión y recuperación, en el que se da un equilibrio entre ambas métricas.

Métricas para Problemas de Regresión:

- **Error Cuadrático Medio (MSE):** Calcula el promedio de los cuadrados de las diferencias entre las predicciones y los valores reales.
- **Raíz del Error Cuadrático Medio (RMSE):** Es la raíz cuadrada del MSE y tiene la misma unidad que la variable objetivo.
- **Error Absoluto Medio (MAE):** Calcula el promedio de las diferencias absolutas entre las predicciones y los valores reales.
- **Coeficiente de Determinación (R^2):** Indica la proporción de la varianza en la variable dependiente que es predecible a partir de la variable independiente.

4 Metodología

En esta investigación la elección de la técnica de clasificación correcta es fundamental para la obtención de resultados óptimos, es por ello que uno de los objetivos principales del trabajo se centra en la revisión y análisis de los diferentes tipos de aprendizaje desarrollados para Machine Learning, como son los árboles de decisión, clustering, deep-learning (redes neuronales), de tal manera que al examinar, desarrollar y programar cada uno de estos métodos, se podrá identificar cuál presenta los mejores resultados en concordancia con la información de entrada.

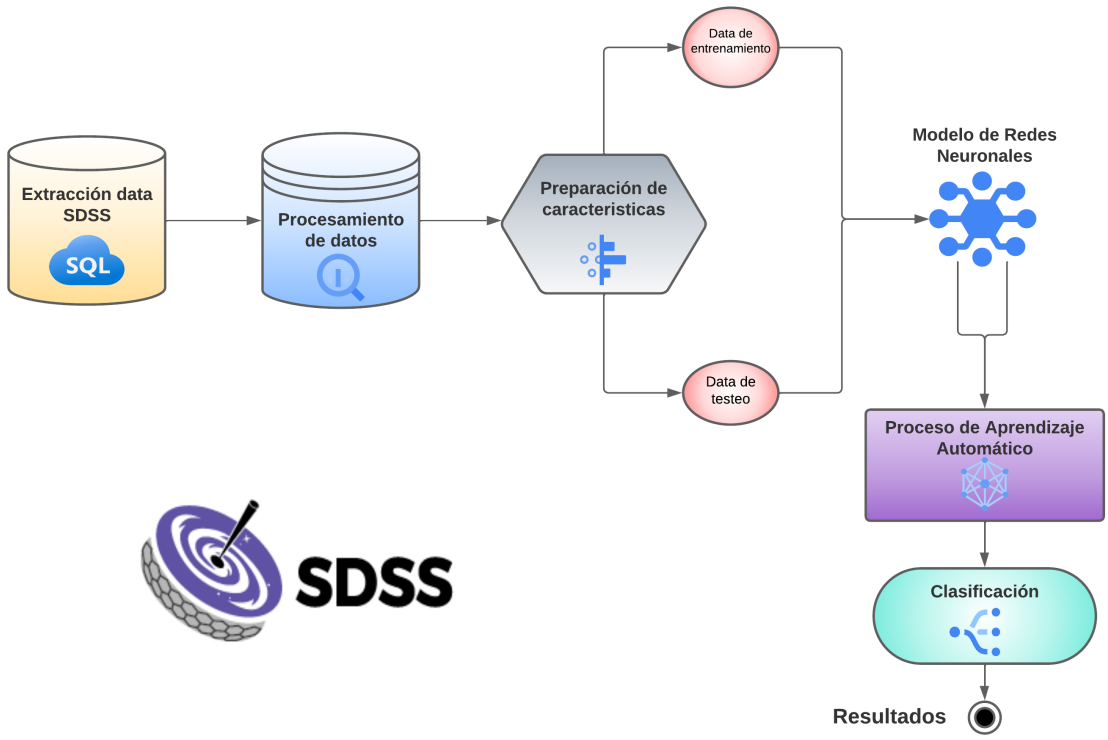


Figura 4-1: Diagrama de bloques que describe el proceso en el algoritmo de Machine Learning.

4.1 Algoritmo Machine Learning

El algoritmo de aprendizaje automático (Machine Learning), desarrollado en esta investigación usando Python[1], consigue, no solo replicar la información obtenida de los diferentes diagramas diagnóstico, sino que estructura una red neuronal capaz de clasificar galaxias activas a diferentes rangos de redshift más allá del Universo Local ($z < 1$), utilizando como parámetros de entrada cada una de las características fotométricas y espectroscópicas de los objetos. El diagrama de la **Figura 4-1** representa el proceso secuencial utilizado en el algoritmo, donde se lleva a cabo una serie de pasos que estructuran al modelo.

4.1.1 Extracción de Datos

Inicialmente, se toma una muestra de 250.000 galaxias de la base de datos del Sloan Digital Sky Survey[2]. A través de la consulta en SQL, se extraen los datos necesarios, tomando la información de las distintas tablas que lo estructuran, estas tres tablas principales son ''galSpecExtra'', ''galSpecLine'' y ''PhotoObj'', las cuales contienen las columnas con la información requerida: ID, Hα, Hβ, [NII], [SII], [OI], [OII], [OIII], u, g, r, i, z, $W_{H\alpha}$ y redshift. En el Apéndice 1 (Tabla A.1) se visualiza la extracción de la base de datos del SDSS con cada una de las columnas anteriormente mencionadas.

4.1.2 Procesamiento de Datos

El segundo bloque se encarga de hacer la limpieza y tratamiento de la información, tomando, inicialmente, todos los objetos con un valor de redshift $z < 1$, filtrando los objetos para tomar solo las galaxias o QSOs, además de eliminar duplicados debido a que varias galaxias tienen el mismo identificador ''SpecObjId'' pero diferentes datos, entonces se verifica aquellos datos negativos o nulos para ser depurados.

4.1.3 Preparación de Características

En el tercer paso se aplica la preparación de los datos para las líneas de emisión y filtros fotométricos, basado en los cocientes de líneas de los diagramas diagnóstico. En la **Figura 2-7** se determina el cálculo de cada una de las características que estructurarán el cubo de datos que alimentará la red neuronal, obteniendo las siguientes columnas: [NII]/Hα, [SII]/Hα,

[1]Ver código Python: `https://github.com/katthe/Machine-Learning-Classification.git`
[2]Ver página web: `https://skyserver.sdss.org/dr17/SearchTools/sql`

[OI]/Hα, [OII]/Hα, [OII]/Hβ, [OIII]/Hβ, U-B [*Blanton & Roweis*. 2007], u-r, g-z, $W_{H\alpha}$. En el Apéndice 2 (Tabla B.1) se estructura el dataframe.

A partir del dataframe generado, es necesario sustraer los datos para entrenamiento, diferentes de la muestra original y que corresponden, aproximadamente, al 10 % de la muestra total (20.000 galaxias aleatorias), la cual servirá al modelo para realizar el proceso de aprendizaje automático a través de las redes neuronales.

La matriz de la **Figura 4-2** muestra las correlaciones entre los distintos datos, en este caso las características de entrada del modelo. Esta medida estadística describe la relación entre dos variables y proporciona una visión general de cómo todas las variables en este conjunto de datos están relacionadas entre sí. Se percibe que algunas variables tienen una fuerte correlación, mientras que para otras su valor es muy bajo, sin embargo al trabajar conjuntamente son trazadores precisos de actividad galáctica.

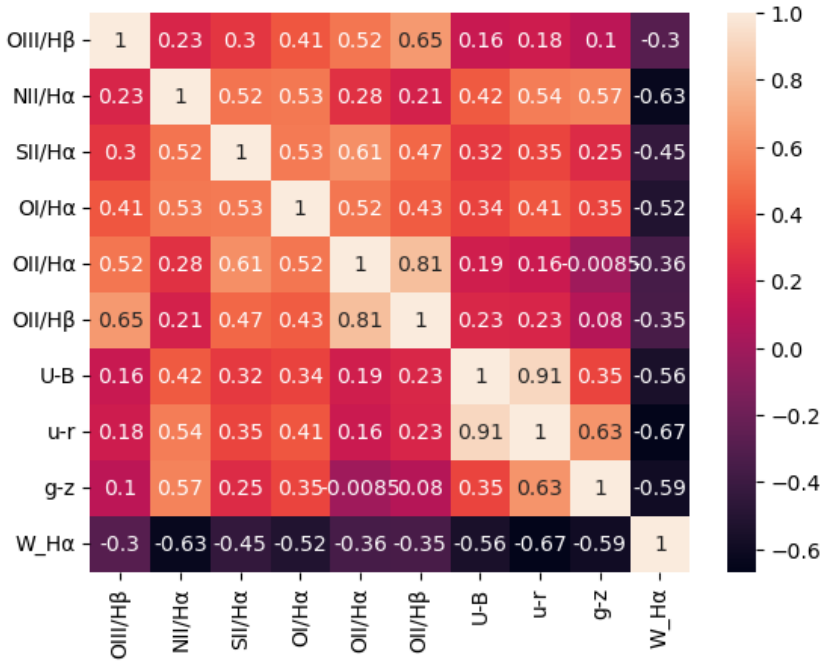


Figura 4-2: Matriz de correlación para las diez características de entrada del modelo, la barra de color y los valores indican el nivel de correlación entre variables, siendo 1 la máxima y 0 la mínima.

4.1.4 Modelo de Redes Neuronales

Como se visualiza en la **Figura 4-3**, la estructura de la red neuronal de aprendizaje profundo, más conocido como Deep Learning, se compone de una serie de neuronas interconectadas entre sí, en cinco capas diferentes, cada una parametrizada con sus respectivas funciones de activación e hiperparámetros.

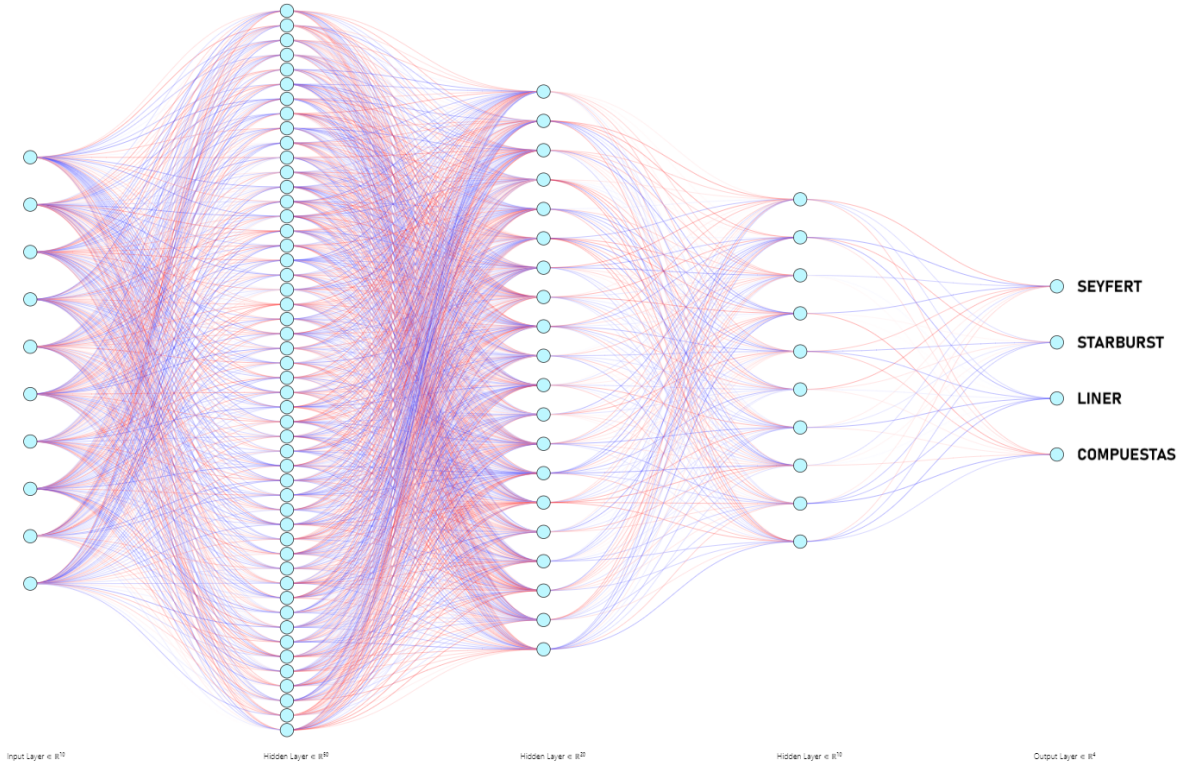


Figura 4-3: Estructura red neuronal diseñada en https://alexlenail.me/NN-SVG/index.html

Modelo Secuencial

En la **Figura 4-4** está representada la arquitectura de esta red neuronal que utiliza un modelo secuencial, una primera capa compuesta por diez entradas correspondientes a las características obtenidas en el dataframe de extracción, una segunda capa con cincuenta neuronas que utilizan una función de activación *''Leaky Relu''*, una función óptima para problemas de clasificación, ya que evita que se desactiven neuronas cuyos valores sean negativos. Para esta función se determina una tasa de aprendizaje $\alpha = 0{,}1$, y un Kernel Inicializador *''He Normal''*, el cual establece los valores iniciales de los pesos de la capa antes de que comience el proceso de entrenamiento.

En la **Figura 4-5** se representa, a través de un histograma, la distribución de pesos con la que se inicializa la red neuronal y los pesos con los que finalizan en la última capa. La media de los pesos debe generalmente tener un valor de cero y, para nuestros datos, se obtiene un $\mu \approx$-0.005 para la primera capa y un $\mu \approx$0.012 para la última capa, su desviación estándar σ que depende del número de entradas de la capa. Se observa como los límites de la gaussiana disminuyen y son mas cercanos a cero, entre la primera y la última iteración, lo que evidencia una óptima evolución en el proceso de aprendizaje.

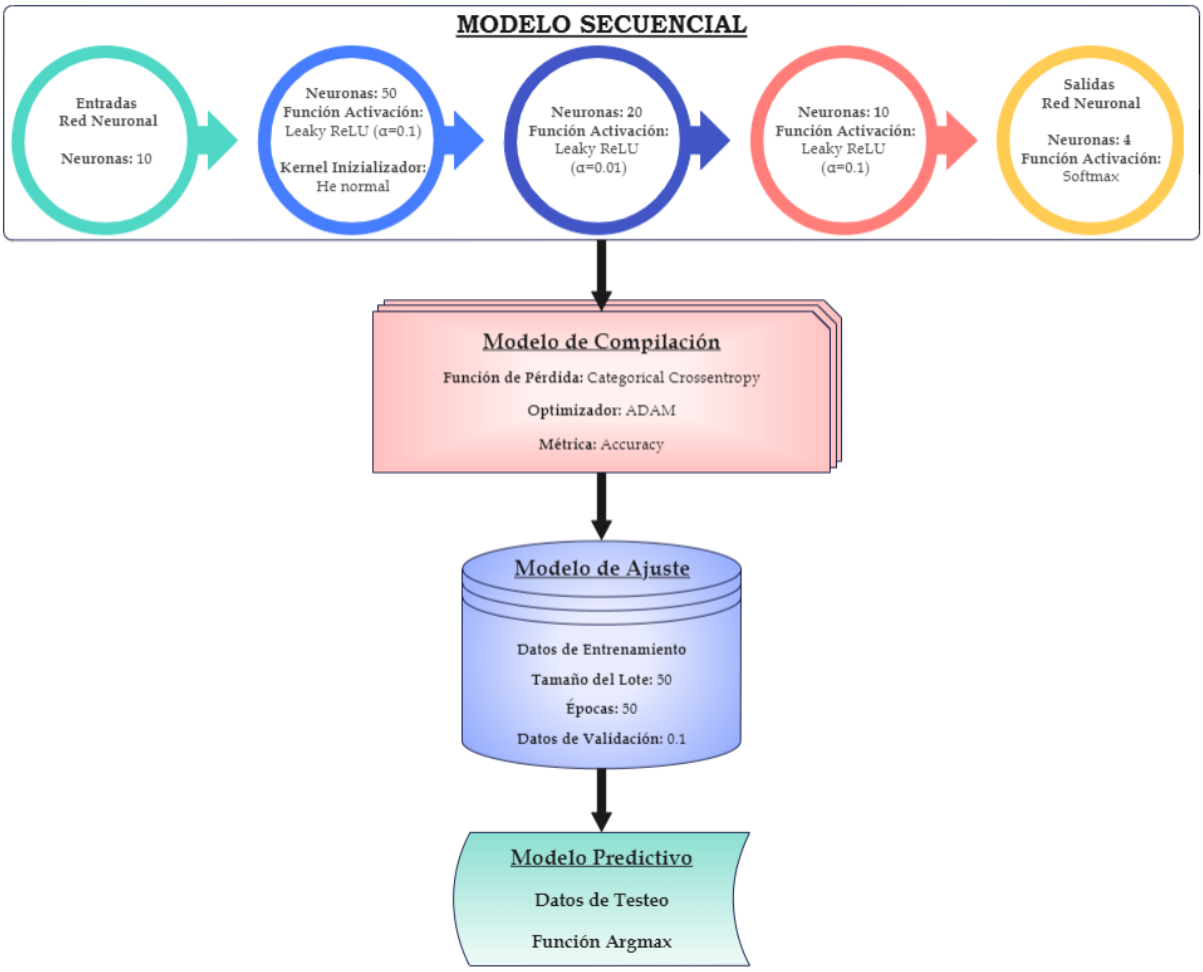


Figura 4-4: Arquitectura interna de la red neuronal representada en diagrama de bloques.

La siguiente capa se compone de veinte neuronas, una función de activación *''Leaky Relu''* y una tasa de aprendizaje $\alpha = 0{,}01$, una cuarta capa de diez neuronas con una función de activación *''Leaky Relu''* igualmente a un $\alpha = 0{,}1$ y, finalmente, la última capa se compone de las cuatro salidas, cuya función de activación es *''Softmax''*, encargada de transformar las salidas en una representación en forma de probabilidades entre 0 y 1, que corresponden al tipo de galaxia etiquetado, ya sea Seyfert, Starburst, LINER o Compuestas.

La arquitectura de la red neuronal está basada en un método de prueba y error que se realizó con múltiples configuraciones, verificando en cada una de ellas cuáles son los parámetros óptimos que dieran como resultado el menor error posible.

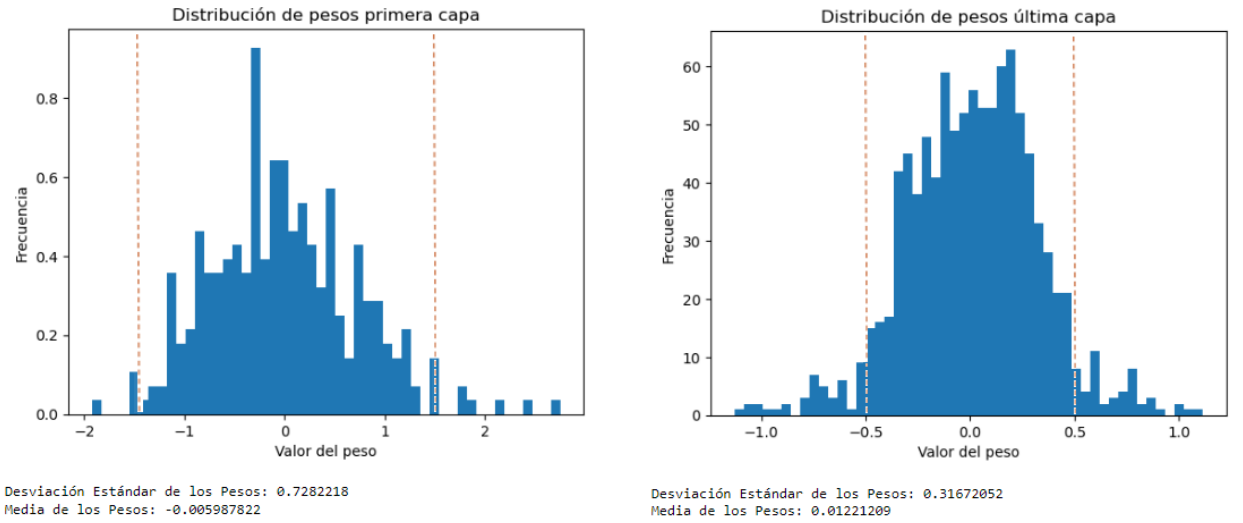


Figura 4-5: Histograma de distribución de pesos, la gráfica izquierda representa la distribución de la primera capa, la gráfica de la derecha representa la distribución en la última capa de la red neuronal

Modelo de Compilación

Método utilizado para estructurar la fase de entrenamiento del modelo de aprendizaje automático. La compilación del modelo implica la configuración de varios aspectos clave que son necesarios para el proceso de entrenamiento. Algunos de los parámetros que se especifican durante la compilación incluyen:

- **Función de pérdida (Categorical Crossentropy):** Especifica cómo se calcula la diferencia entre las predicciones del modelo y las etiquetas verdaderas. El objetivo durante el entrenamiento es minimizar esta función.
- **Optimizador (Adaptive Moment Estimation):** Algoritmo de optimización de gradiente estocástico que verifica la evolución del *Learning Rate* en cada iteración para encontrar el punto mínimo del error del entrenamiento.
- **Métrica (Accuracy):** Esta métrica de precisión es una medida utilizada para evaluar el rendimiento del modelo de clasificación. La precisión se define como la fracción de instancias correctamente clasificadas sobre el total de instancias.

Modelo de Ajuste

Este método toma como entrada los datos de entrenamiento previamente obtenidos (características y etiquetas), para realizar el proceso de optimización que ajusta los parámetros del modelo y devuelve información sobre la evolución del entrenamiento.

- **Datos en X (x train):** Datos o características de entrenamiento.
- **Datos en Y (y train):** Etiquetas correspondientes a los datos de entrenamiento.
- **Épocas (Epochs=50):** Número de épocas o iteraciones completas a través de todo el conjunto de entrenamiento.
- **Tamaño del lote (Batch Size=50):** Especifica cuántas muestras se utilizan para actualizar los pesos del modelo en cada paso de optimización.
- **Datos de validación (Validation Split=0.1):** Datos y etiquetas del conjunto de validación que se utilizan para evaluar el rendimiento del modelo en datos no vistos durante el entrenamiento, en este caso el 10 % de los datos corresponde a 2000 galaxias.

Durante el proceso de entrenamiento, el modelo ajusta sus parámetros (pesos y sesgos) utilizando el algoritmo de optimización especificado durante la compilación del modelo. Este proceso implica propagar hacia atrás el error calculado entre las predicciones del modelo y las etiquetas verdaderas, y luego actualizar los pesos del modelo en la dirección que minimiza este error.

El método fit devuelve un objeto ‘‘History’’ que contiene información sobre las métricas de entrenamiento y validación en cada época. Esto se utiliza para visualizar y analizar el rendimiento del modelo a lo largo del tiempo (**Figura 4-8**).

Modelo Predictivo

Los datos de prueba o testeo entran a este método que se utiliza para realizar predicciones, obteniendo como resultado la clasificación para cada una de las galaxias de la muestra a partir del aprendizaje adquirido en el entrenamiento.

1. **Entrada de datos:** Proporciona el conjunto de datos de entrada (x) para el cual se realiza la predicción. Este conjunto de datos debe tener la misma estructura que los datos de entrenamiento que se utilizaron para entrenar el modelo.
2. **Propagación hacia Adelante (Forward Propagation):** Los datos de entrada se pasan a través del modelo capa por capa, utilizando el proceso de propagación hacia adelante. Cada capa realiza operaciones específicas, como multiplicaciones de matrices seguidas de aplicaciones de funciones de activación.
3. **Predicciones:** Se obtienen como salida de la última capa del modelo. La salida es una distribución de probabilidad, donde se utiliza la función ‘‘argmax’’ para encontrar la clase con la probabilidad más alta y así determinar la clase predicha.

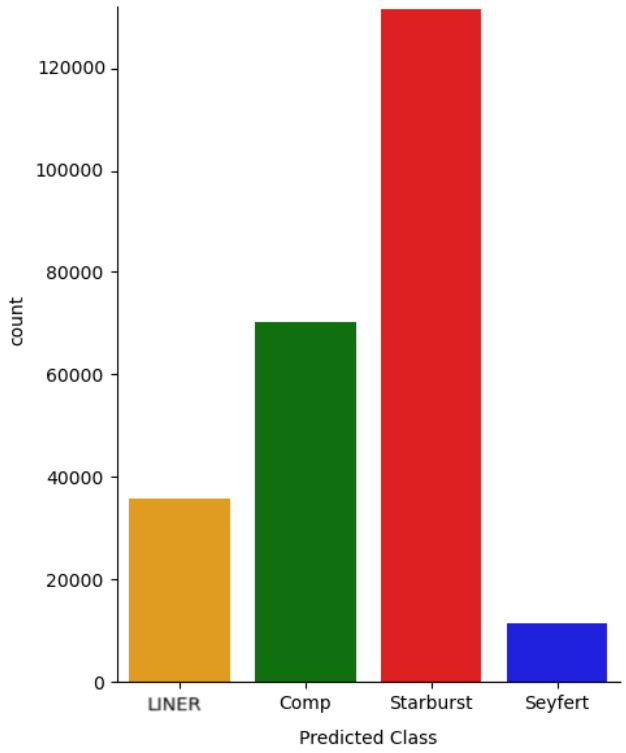


Figura 4-6: Utilizando la notación de colores las galaxias Seyfert pertenecen al color rojo, las Compuestas al verde, las LINER al amarillo y las Seyfert al color azul.

La **Figura 4-6** grafica en un histograma los 4 grupos de galaxias y la cantidad de objetos pertenecientes a cada una de estas clases, esto permite verificar la distribución total de la muestra.

Para obtener la precisión con la que está trabajando el modelo se emplean varias herramientas, una de ellas es la matriz de confusión, que permite visualizar el desempeño de un algoritmo de aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real, permitiendo ver qué tipos de aciertos y errores está teniendo el modelo a la hora de pasar por el proceso de aprendizaje con los datos.

La **Figura 4-7** muestra en su diagonal la cantidad de galaxias que se predijeron correctamente por cada clase, mientras que los demás datos de la matriz nos indica los falsos positivos, evidenciando una alta población de predicciones.

El modelo de la **Figura 4-8** representa la evolución del algoritmo de aprendizaje por cada una de la épocas, en este caso 50 en total, reflejando cómo los datos de validación en cada ciclo disminuyen su diferencia respecto a los datos de entrenamiento, dando como resultado un porcentaje de precisión del 99.27 % y un error de tan solo el 0.72 %.

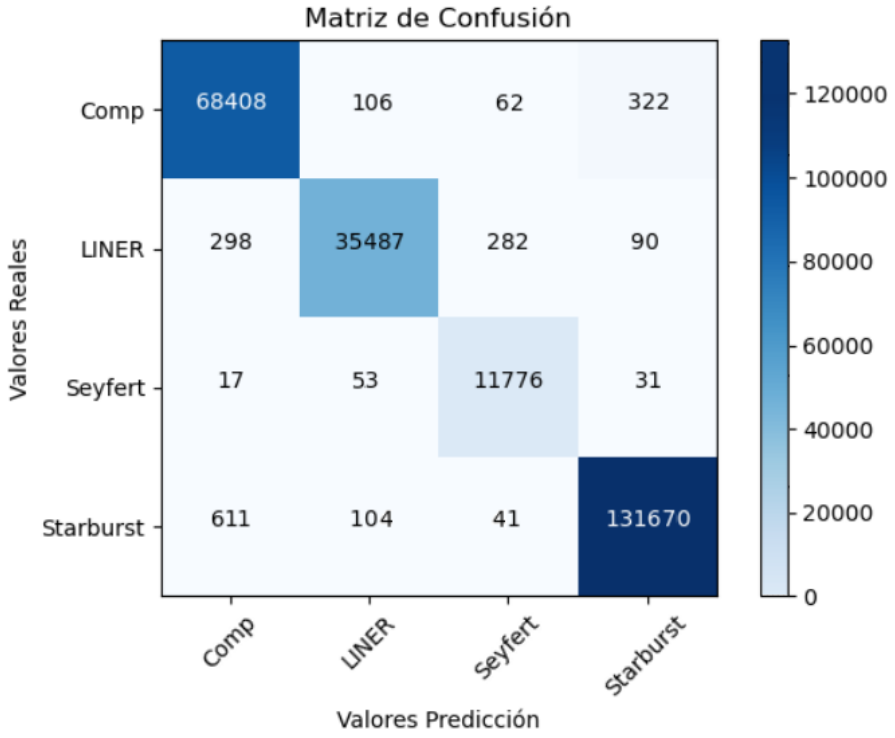


Figura 4-7: Matriz de confusión para los cuatro tipos de galaxias, la barra de color representa los niveles de predicción de cada clasificación de verdaderos positivos a falsos positivos.

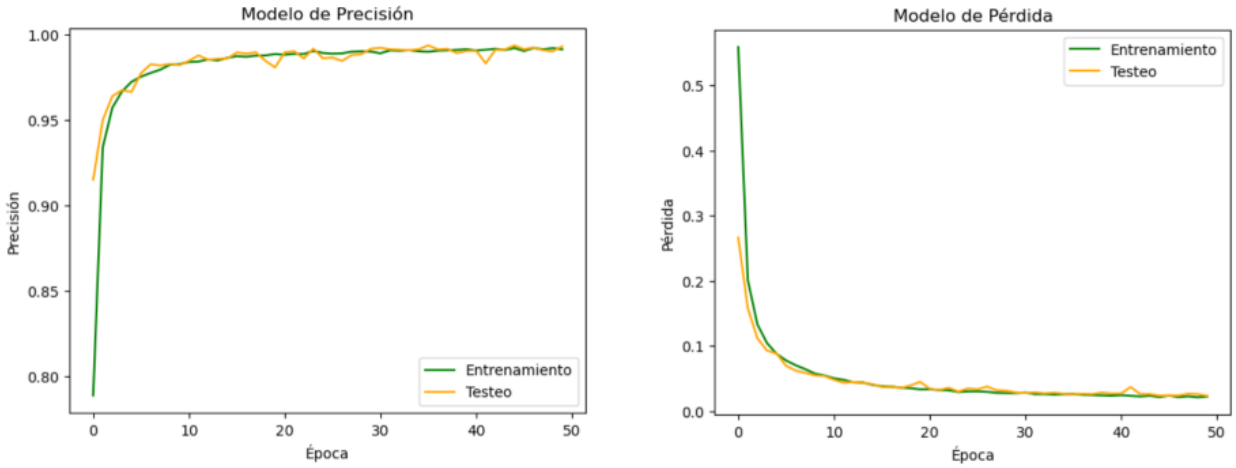


Figura 4-8: Gráfica para el modelo de eficiencia del algoritmo, a la izquierda el modelo de precisión para los datos de testeo y de entrenamiento, a la derecha representación del modelo de pérdida para los datos de testeo y de entrenamiento.

4.2 Modelo de Aprendizaje Supervisado (Árbol de Decisión)

En problemas de clasificación, los árboles de decisión son una opción que logra determinar, a través de aprendizaje supervisado, los grupos de galaxias en los que se divide la muestra en estudio.

Para poder evaluar la información utilizando este modelo de aprendizaje se realizan los siguientes pasos:

- Inicialmente, se importa de la librería Scikit-Learn la clase *''train_test_split''*, la cual toma dos argumentos, el primero es el dataframe con las características de entrada, en este caso los cocientes de las líneas de emisión, el segundo argumento es el tamaño de la muestra de entrenamiento que será igual a 0.6 es decir el 60 % del total de la muestra, de esta manera se obtiene el cubo de datos tanto de entrenamiento como de testeo.
- En segunda instancia se utiliza un clasificador de árbol de decisión llamado *''DecisionTreeClassifier''* donde se define la profundidad máxima de capas que tendrá el árbol, en este caso un total de 10.
- Para el entrenamiento del algoritmo se implementa el modelo de ajuste, que toma tanto los datos de entrenamiento como los de testeo. El algoritmo comienza seleccionando la característica que mejor separa los datos en términos de la variable objetivo, en este caso la etiqueta de clasificación de las galaxias, esto se hace utilizando la métrica impureza de Gini, que se define como la probabilidad de clasificar incorrectamente un punto de datos aleatorio en el conjunto de datos si se etiquetara en función de la distribución de clases del conjunto de datos, si el conjunto pertenece a una clase, entonces su impureza es cero, tal como lo muestra el último nivel del árbol de decisión de la **Figura 4-9 b)**.
- Después de seleccionar la característica, referente a los cocientes de líneas, el conjunto de datos se divide en subconjuntos más pequeños en función de los valores de esa característica. Este proceso se repite recursivamente para cada subconjunto, hasta que se cumple un criterio de parada, dado por la profundidad máxima del árbol.
- Cuando se alcanza un nodo que cumple el criterio de parada, se convierte en un nodo hoja y se asigna la etiqueta de clase, en este caso la clasificación de la galaxia, llegando así finalmente a la predicción.
- Se obtienen los resultados del modelo y se grafica el árbol de decisión como se visualiza en la **Figura 4-9 a)**, donde se muestra la arquitectura general a partir de un nodo raíz y las diferentes ramas que generan la estructura, completando 10 niveles de profundidad.
- La **Figura 4-9 b)** muestra un acercamiento de las hojas finales del árbol de decisión, donde se presenta la predicción del modelo con su correspondiente etiqueta del tipo de galaxia, ya sea Starburst, Seyfert, LINER o Compuesta.

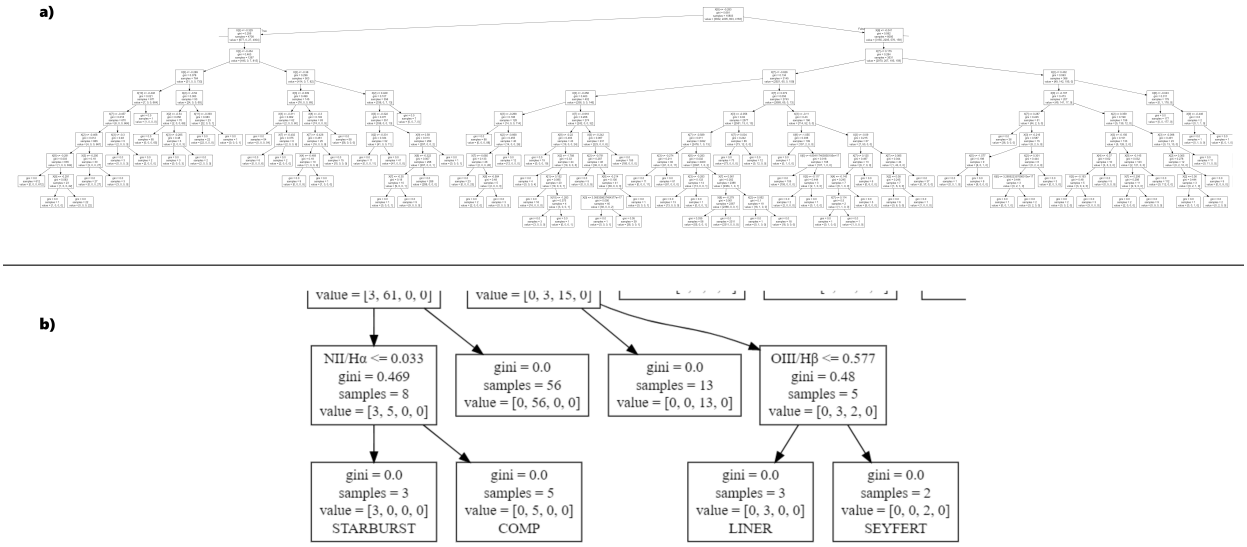


Figura 4-9: a) Arquitectura del árbol de decisión, nodo raíz, nodos de decisión y hojas. b) Amplificación de los nodos finales del árbol de decisión, con los valores de las características, la métrica, las muestras y la clasificación. Diagrama generado en https://dreampuf.github.io/GraphvizOnline

4.2.1 Eficiencia

La **Figura 4-10** representa la evolución del algoritmo de aprendizaje por cada una de la capas de profundidad del árbol de decisión, en este caso se define un límite de 20 capas, se refleja cómo, a partir de la capa número 10, los datos de entrenamiento llegan a un 100 % y la data de testeo se acerca a un 97 % de precisión, por lo tanto se obtiene un 3 % de error. La razón por la cual se ha elegido que los niveles de profundidad del árbol sea igual a 10 se debe a que en este punto el modelo se estabiliza respecto a la máxima eficiencia, evitando así un sobreajuste en los datos.

4.3 Modelo de Aprendizaje No Supervisado (Clustering)

Dentro de los múltiples modelos desarrollados para clasificación en Machine Learning, uno de los más populares en el aprendizaje no supervisado es el llamado Clustering o Agrupamiento, cuyo objetivo es el de identificar y reconocer patrones dentro de la muestra de datos de entrada, y así lograr agruparlos según sus características. El proceso para poder ejecutar este algoritmo se realiza siguiendo los parámetros a continuación:

- El primer paso es importar la clase *''KMeans''* de la librería Scikit-Learn, al crear este objeto se debe indicar como parámetro el número de clústeres deseado, en este caso 4,

```
>1, train: 0.775, test: 0.775
>2, train: 0.859, test: 0.859
>3, train: 0.915, test: 0.898
>4, train: 0.946, test: 0.935
>5, train: 0.966, test: 0.952
>6, train: 0.982, test: 0.962
>7, train: 0.994, test: 0.964
>8, train: 0.997, test: 0.972
>9, train: 0.999, test: 0.966
>10, train: 0.999, test: 0.970
>11, train: 1.000, test: 0.968
>12, train: 1.000, test: 0.973
>13, train: 1.000, test: 0.971
>14, train: 1.000, test: 0.967
>15, train: 1.000, test: 0.966
>16, train: 1.000, test: 0.971
>17, train: 1.000, test: 0.971
>18, train: 1.000, test: 0.967
>19, train: 1.000, test: 0.972
>20, train: 1.000, test: 0.972
```

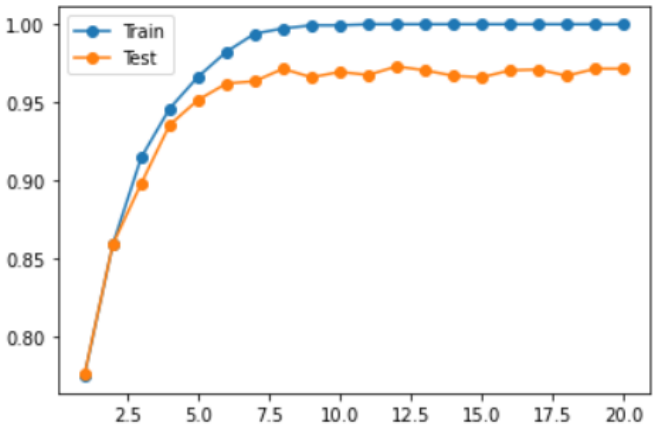


Figura 4-10: Gráfica para el modelo de eficiencia del árbol de decisión para los datos de testeo y de entrenamiento

que hace referencia a la cantidad de clases por clasificar de los datos.

- El modelo KMeans se entrena con los datos mediante el método *"fit"*. Durante el proceso de entrenamiento, el algoritmo asigna puntos de datos a clústeres y ajusta los centroides de manera iterativa para minimizar la suma de las distancias cuadráticas entre los puntos de datos y los centroides de sus clústeres asignados. El proceso se divide en 4 pasos:

  1. **Inicialización de centroides:** El algoritmo comienza seleccionando aleatoriamente los 4 puntos de datos como centroides iniciales.
  2. **Asignación de puntos a clústeres:** Para cada punto de datos en la matriz de entrada, el algoritmo calcula la distancia euclidiana cuadrada entre el punto y todos los centroides, el punto se asigna al clúster cuyo centroide tiene la distancia más pequeña.
  3. **Actualización de centroides:** Después de asignar todos los puntos a clústeres, los centroides se recalculan como el promedio de las coordenadas de los puntos asignados a cada clúster.
  4. **Repetición:** Los pasos 2 y 3 se repiten iterativamente hasta que se alcanza la convergencia, es decir, cuando los centroides ya no cambian significativamente entre iteraciones o se alcanza el número máximo de iteraciones especificado por el parámetro *"max_iter"*.

- Una vez se ha realizado el entrenamiento, el método *"predict"* se utiliza para asignar nuevos puntos de datos a los clústeres aprendidos por el modelo previamente. El método devuelve un array de enteros que representan las etiquetas de clúster asignadas a cada punto de datos en la matriz de entrada, cada valor corresponde al índice del clúster al que se ha asignado el punto.

- Finalizado el proceso, se obtienen las etiquetas de clúster para cada punto de datos, utilizando el comando *"kmeans.labels_"* y las coordenadas de los centroides con *"kmeans.cluster_centers_"*

La visualización gráfica de la **Figura 5-4** permite revisar y analizar el proceso de entrenamiento que ejecutó este tipo de aprendizaje no supervisado y así, de esta manera verificar si los resultados son congruentes con el desarrollo que se está realizando.

5 Resultados

Una de las razones principales por la que se desarrolló esta investigación se debe a que, en las diferentes bases de datos públicas consultadas, como la del Sloan Digital Sky Survey (SDSS), Infrared Astronomical Satellite (IRAS), NASA/IPAC Extragalactic Database (NED), SIMBAD Astronomical Database - CDS, la información que tienen referente a la clasificación de galaxias en imprecisa y errónea.

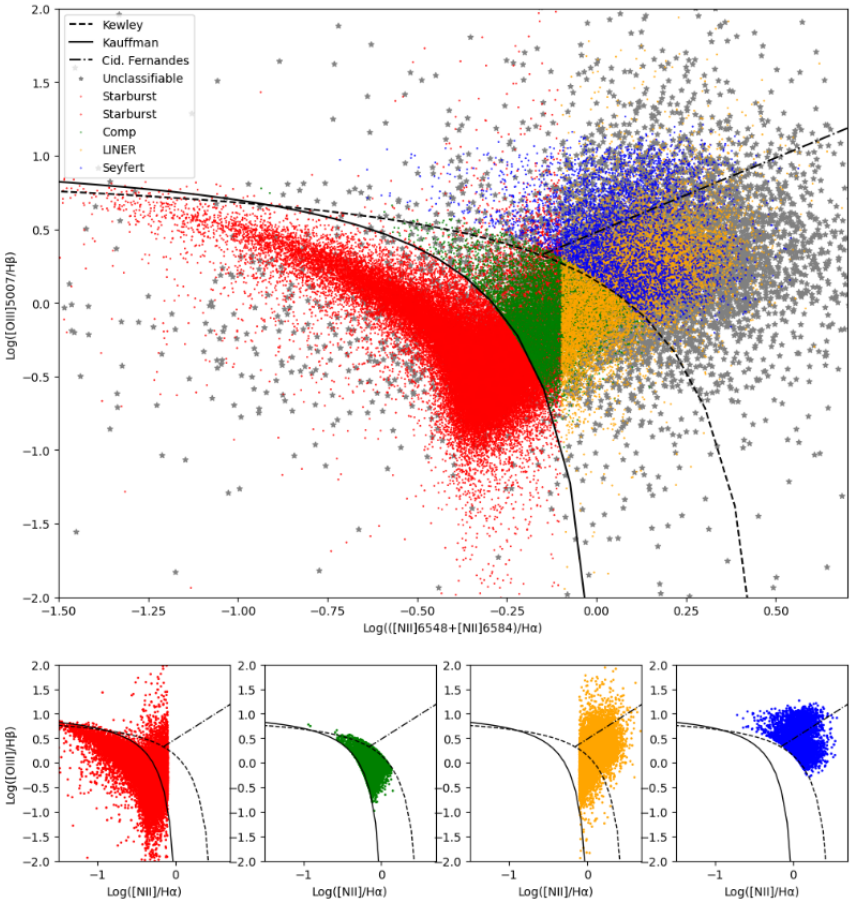


Figura 5-1: Clasificación de galaxias en diagrama BPT tomando información de la base de datos SDSS, representado en Python.

CLASIFICACIÓN	SDSS	Machine Learning
Starburst	159661	131670
Seyfert	12312	11776
LINER	29931	35487
Compuestas	24009	68408
No Clasificables	24087	------

Tabla 5-1: Clasificación de galaxias calculadas por la base de datos SDSS vs. Clasificación del algoritmo de Machine Learning

Un ejemplo de ello se expone en la **Figura 5-1**, donde se llevó a cabo la extracción de datos de la base del SDSS. Esta extracción incluyó una columna denominada *''bptclass''*, la cual refleja la clasificación de las líneas de emisión según el diagrama BPT, siguiendo la metodología descrita por [*Brinchmann et al.*. 2004]. En esta clasificación, el valor -1 indica que es no clasificable, 1 para formación de estrellas, 2 señala formación de estrellas con baja relación S/N, 3 corresponde a galaxias compuestas, 4 a AGNs y 5 representa las tipo LINER.

En los subpaneles de la gráfica, se destaca cómo un considerable número de galaxias no se ajusta a los límites que definen cada categoría de clasificación. Tal es el caso de las galaxias Starburst, en color rojo, que aparecen dispersas entre las zonas designadas para galaxias compuestas y AGNs. Similarmente, ocurre con las Seyfert y LINER, además de esto existe otro grupo de galaxias identificadas como no clasificables, marcadas en color gris. Este hallazgo sugiere que las etiquetas proporcionadas directamente en la base de datos pueden carecer de información confiable.

La **Tabla 5-1** revela el número de galaxias clasificadas para cada grupo, dentro del SDSS, en comparación con la clasificación final que arrojó el algoritmo de Machine Learning. Se constata, efectivamente, que las etiquetas de la base de datos es incorrecta.

5.1 Árbol de Decisión

El primer método de aprendizaje utilizado en este trabajo fue el aprendizaje supervisado, donde se escogió el modelo de árbol de decisión, un algoritmo capaz de realizar clasificación múltiple. La clave para encontrar la mejor clasificación se remite al número de niveles de profundidad del árbol, trabajar un valor muy pequeño o muy grande incurre en obtener resultados subestimados o sobreajustados, donde pueden existir falsos positivos.

La **Figura 5-2** reconstruye el diagrama diagnóstico BPT, utilizando un árbol de decisión con una profundidad de 5 capas, se visualiza inicialmente cómo, al ser un valor pequeño, la clasificación es incorrecta bajo las áreas de las curvas delimitadas, encontrando un error de aproximadamente el 5 % según la **Figura 4-10**, lo que representa cerca de 14000 galaxias clasificadas erróneamente.

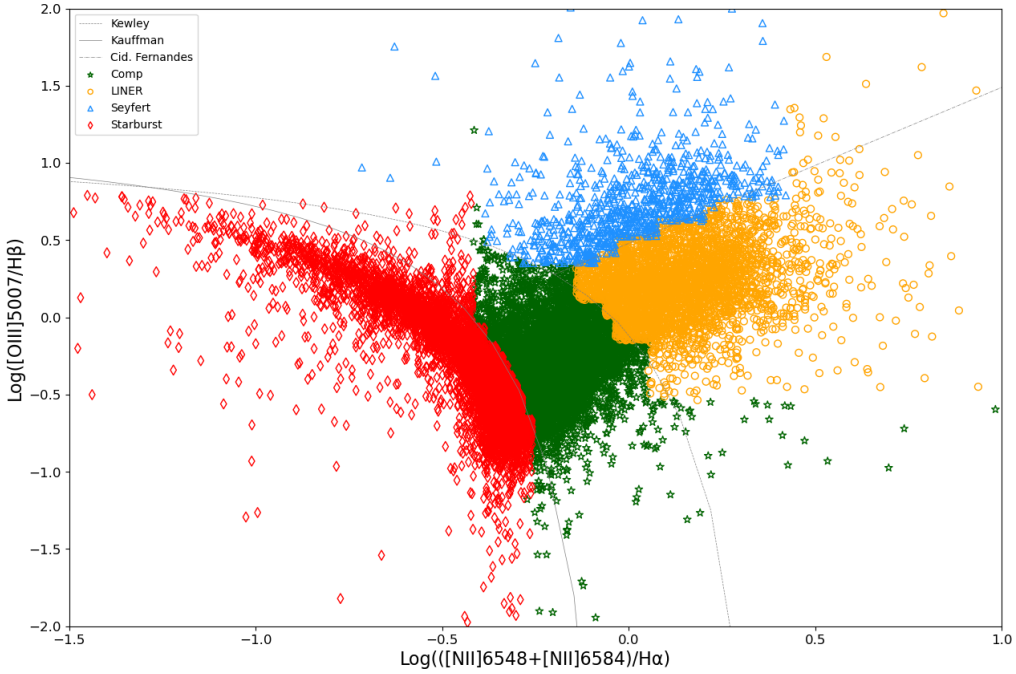


Figura 5-2: Diagrama Diagnóstico BPT para clasificación de galaxias con un modelo de árbol de decisión con 5 capas de profundidad, representado en Python.

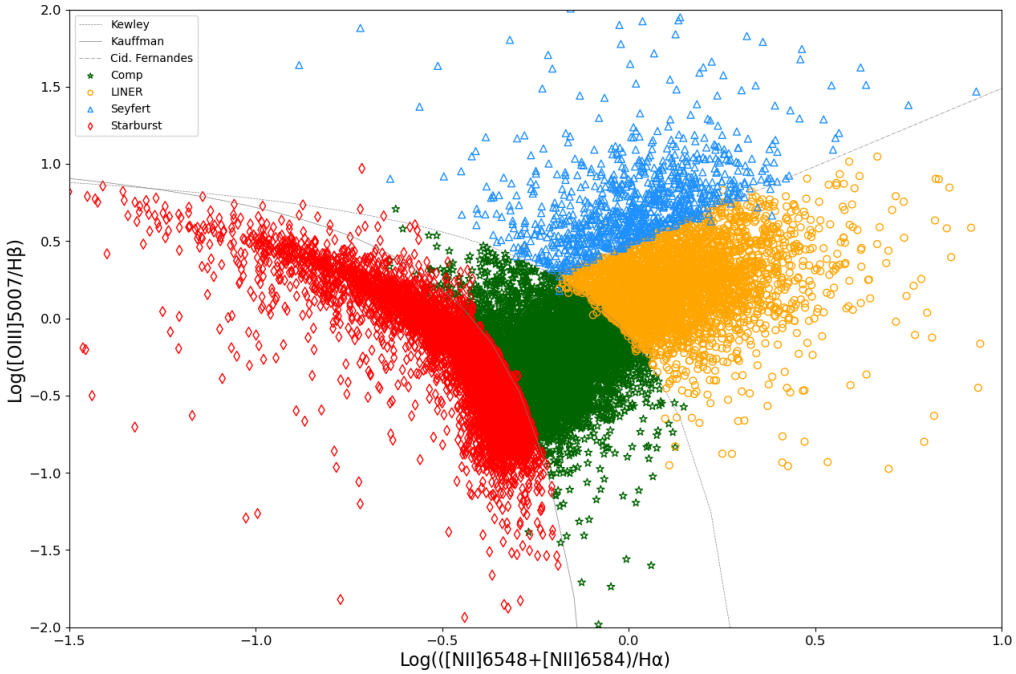


Figura 5-3: Diagrama Diagnóstico BPT para clasificación de galaxias con un modelo de árbol de decisión con 10 capas de profundidad, representado en Python.

Al realizar una nueva ejecución aumentando el número de capas de profundidad a 10, el diagrama de la **Figura 5-3**, aunque muestra mejor agrupamiento de las galaxias respecto a su grupo de clasificación correspondiente, se visualizan algunos datos que sobrepasan los límites de las curvas de división, mostrando una eficiencia que alcanza el 97 %, con un margen de error de unas 8000 galaxias aproximadamente.

Cabe destacar que este algoritmo, aunque encuentra una correcta clasificación de las galaxias, requiere al menos el 60 % de la muestra total para utilizarla como data de entrenamiento del modelo, lo que reduce el número total de galaxias a clasificar.

5.2 Clustering

Uno de los objetivos de este trabajo se centró en examinar los diferentes tipos de aprendizaje y los resultados obtenidos para el modelo de aprendizaje no supervisado mediante el algoritmo K-means se presentan en la **Figura 5-4**.

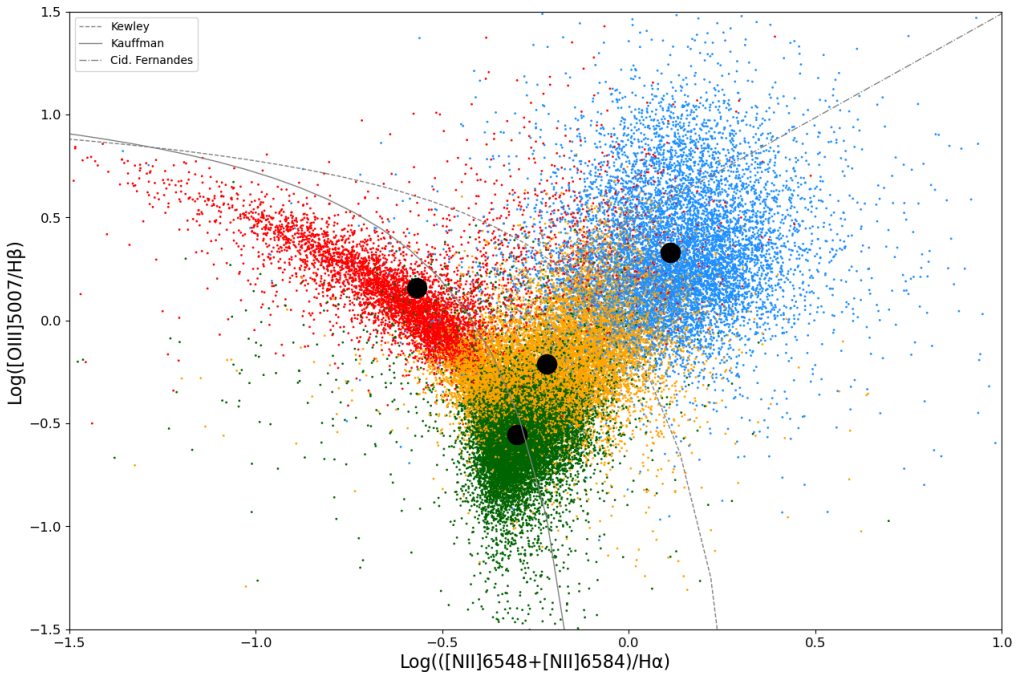


Figura 5-4: Diagrama Diagnóstico BPT para clasificación de galaxias con un modelo de clustering con cuatro centroides, representado en Python

En primera instancia, aunque el algoritmo logra agrupar de manera adecuada cuatro conjuntos distintos de galaxias en un intento por identificar patrones en las características de los datos, se revela insuficiente para la clasificación específica que se busca. Este desafío se atribuye a que las áreas de distribución de los datos están delimitadas por las curvas definidas en el diagrama BPT, lo que genera dificultad en el proceso de entrenamiento.

Dado que este tipo de aprendizaje no utiliza datos de entrenamiento y el modelo carece de una base de comparación para encontrar la posición precisa de los centroides, se concluye que este enfoque no es óptimo para la clasificación de galaxias activas, por lo tanto, se descarta su utilización en el proyecto.

5.3 Deep Learning

Los resultados obtenidos una vez el algoritmo de clasificación ha sido ejecutado, utilizando la metodología de aprendizaje profundo con redes neuronales, se organizan en un cubo de datos que permite hacer la visualización en los diferentes diagramas mostrados a continuación:

5.3.1 Proyección 2D

Diagrama Diagnóstico BPT

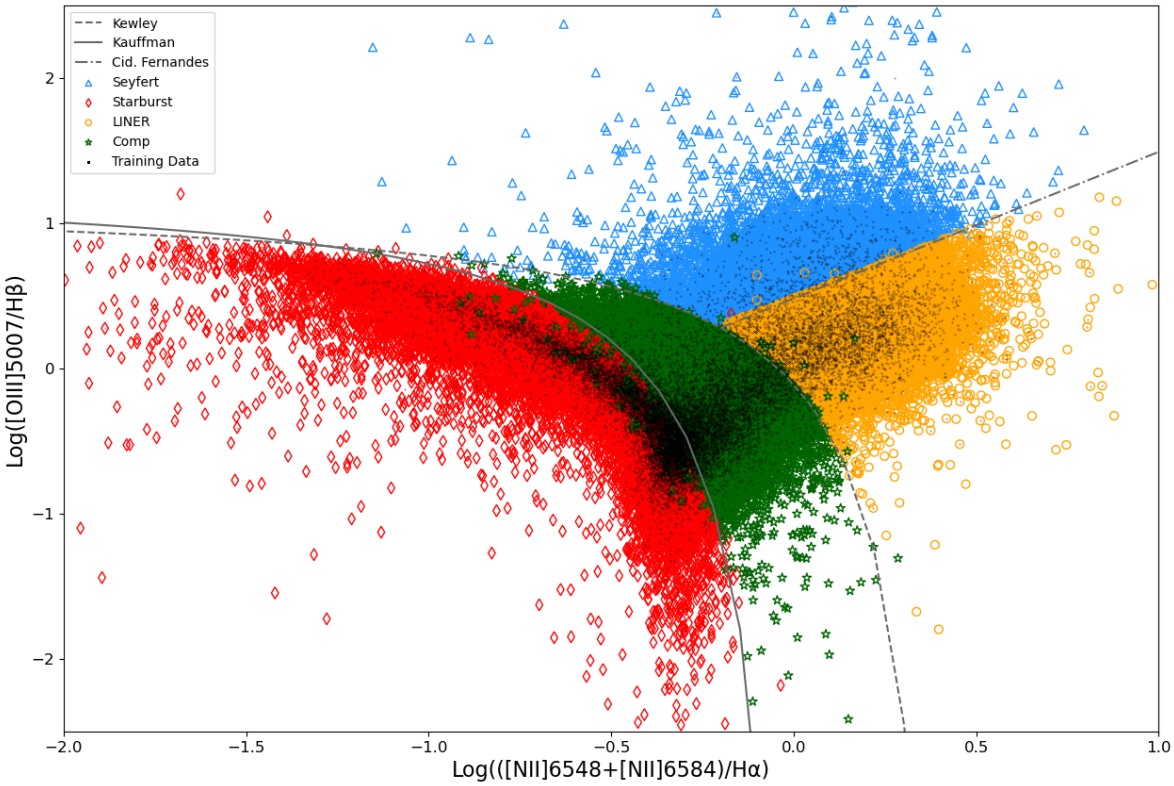


Figura 5-5: Diagrama Diagnóstico BPT para clasificación de galaxias con un modelo de redes neuronales, representado en Python

La **Figura 5-5** reconstruye el diagrama clásico BPT con la clasificación de los cuatro grupos de galaxias, Starburst bajo la curva de Kauffman, Seyfert 2 sobre la línea de Cid Fernandes, Compuestas entre Kauffman y Kewley y las tipo LINER bajo la línea de Cid Fernandes, las galaxias en color negro representan la muestra utilizada para el entrenamiento.

El modelo de precisión de la red neuronal presenta un 99.27 % de eficiencia, según la grafica de la **Figura 4-8**, lo cual se evidencia visualmente en la imagen **5-5**, encontrando una excelente clasificación con un pequeño porcentaje de galaxias que cruza los límites de las curvas teóricas.

El objetivo de utilizar los demás diagramas diagnóstico es el de asegurar la clasificación de aquellos objetos que obtuvieron resultados desacertados en esta primera iteración, para lograr minimizar así el error y asegurar que cada una de las galaxias se etiquete de manera correcta.

Diagrama Diagnóstico WHAN

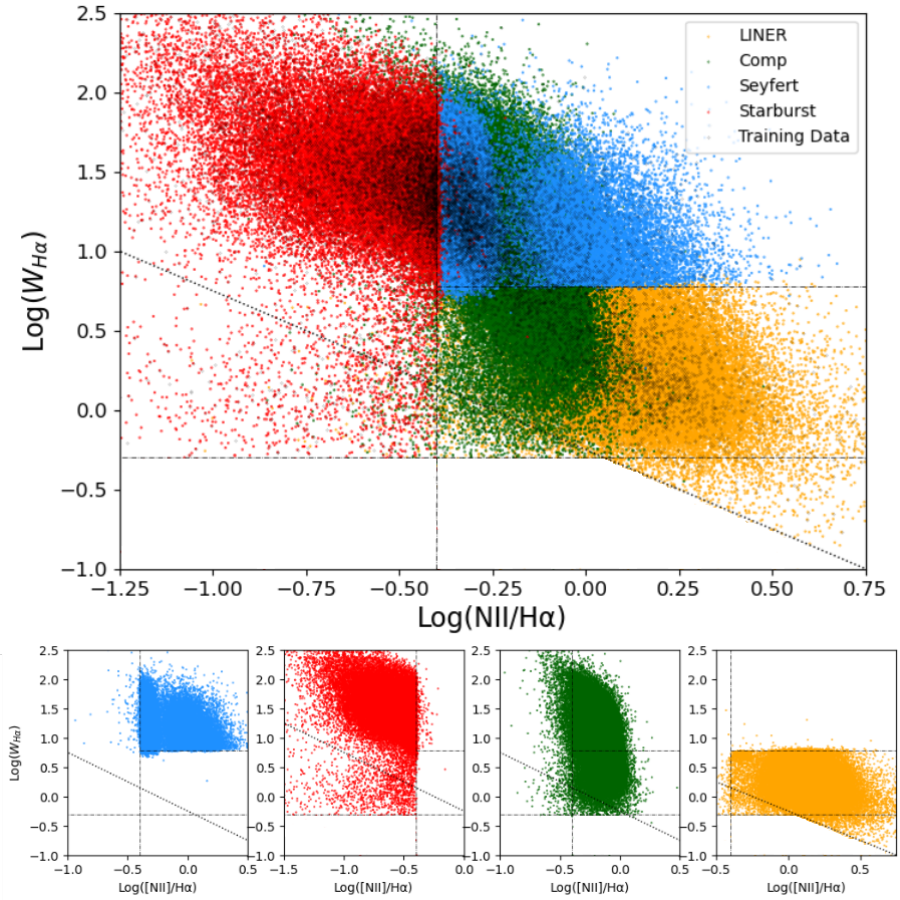


Figura 5-6: Diagrama Diagnóstico WHAN generado a través de Machine Learning, representado en Python.

La **Figura 5-6** muestra los resultados del diagrama WHAN utilizando datos obtenidos de la red neuronal de clasificación, representando la segregación de las cuatro poblaciones en los paneles inferiores. Se observa que las galaxias Seyfert (SY) tienen valores más altos de $W_{H\alpha}$ en comparación con las LINER. Esta diferencia se atribuye al hecho de que las relaciones entre las líneas de emisión trazan las condiciones físicas en el gas ionizado, mientras que $W_{H\alpha}$ mide el poder de la fuente ionizante con respecto a la salida óptica de la población estelar del huésped, derivando así esta distinción entre ambos grupos.

La línea vertical en Log[NII]/Hα = -0,40 corresponde a la transposición óptima de la división entre galaxias Starburst en color rojo y las galaxias Seyfert en color azul (SF/AGN), la línea en $W_{H\alpha}$= 0.75Å representa la división entre galaxias Seyfert y LINER en color amarillo, las galaxias Compuestas se visualizan en color verde y las líneas punteadas que marcan los límites $W_{H\alpha}$ = 0,5Å y Log[NII]/Hα < 0,5, por debajo de los cuales las mediciones de las líneas son inciertas y se muestra el grupo de galaxias retiradas, objetos que son filtrados de la muestra ya que no pertenecen a galaxias activas y se conocen como "Falsos AGNs".

Diagrama Diagnóstico Azul

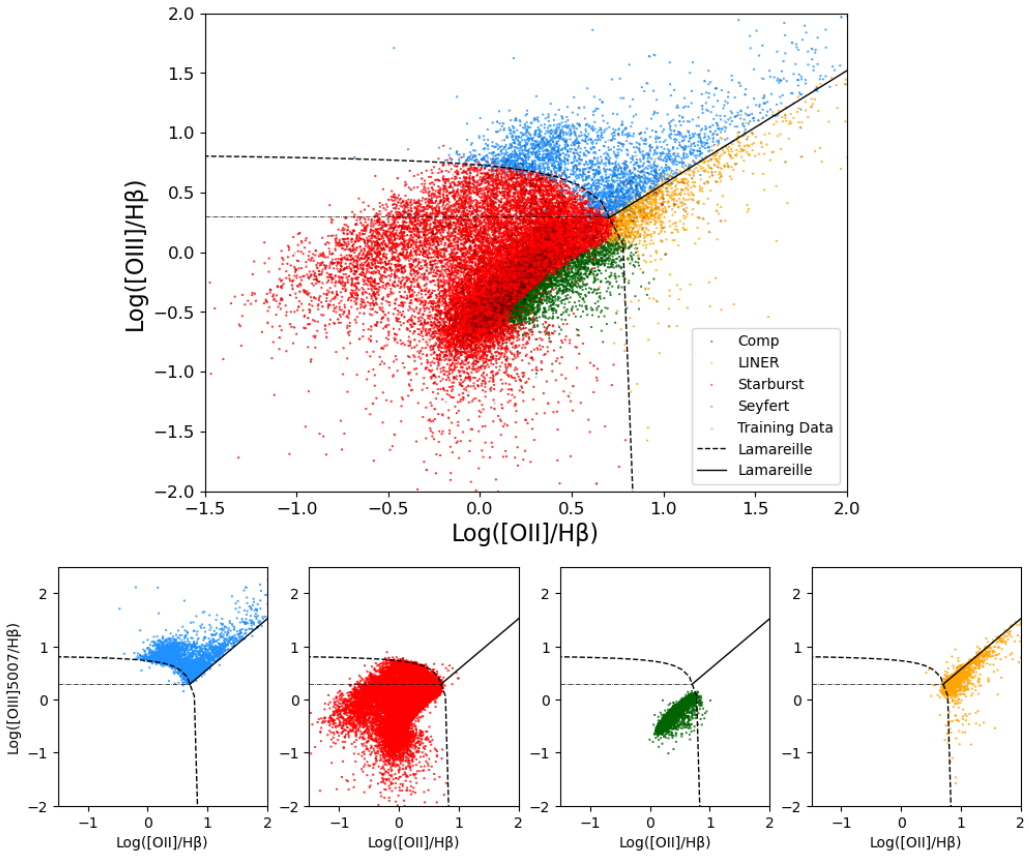


Figura 5-7: Diagrama Diagnóstico Azul generado a través de Machine Learning, representado en Python.

CLASIFICACIÓN	DIAGRAMA BPT	DIAGRAMAS DIAGNÓSTICO
Starburst	9639	16826
Seyfert	3195	2882
LINER	6474	4951
Compuestas	10158	4807

Tabla 5-2: Número de galaxias reclasificadas entre el diagrama diagnóstico BPT y los diagramas diagnóstico Azul y U-B.

En el siguiente diagrama diagnóstico, conocido como el diagrama azul, el enfoque principal reside en la identificación de galaxias con alto corrimiento al rojo $0{,}4 < z < 1$. El objetivo es filtrar esta población específica de objetos para posteriormente reclasificarlos de acuerdo con las curvas predefinidas que delimitan cada categoría. Una vez que se determina a qué grupo pertenecen, se actualiza la etiqueta de cada objeto.

La **Figura 5-7** ilustra la distinción entre el grupo de galaxias Starburst (SF), que se encuentra bajo la curva punteada, y las galaxias tipo Seyfert, que se agrupan sobre la misma curva. También se observa un pequeño conjunto de objetos entre la línea punteada horizontal y la curva, denominado SF-SY, que comprende galaxias con formación estelar y núcleo activo. Las galaxias compuestas se ubican en la zona predeterminada teóricamente y las LINER se posicionan en la parte inferior de la línea recta.

Cabe destacar que este diagrama está diseñado para galaxias con un alto corrimiento al rojo, es por ello que en la **Tabla 5-2** se evidencia el número de galaxias que fueron reclasificadas del tradicional diagrama BPT.

Diagrama Diagnóstico U-B

El índice de color U-B, de igual manera, trabaja en el contexto de galaxias con un alto corrimiento al rojo, reclasificando aquellas que fueron clasificadas incorrectamente según los diagramas tradicionales.

En la **Figura 5-8**, se presentan todos los objetos clasificados por la red neuronal, aquellos situados por debajo de la línea punteada son identificados como galaxias tipo Starburst, mientras que las que se encuentran por encima de esta línea son clasificadas como tipo AGN (Seyfert y LINER). Las galaxias restantes ubicadas en la zona de transición son las llamadas Compuestas.

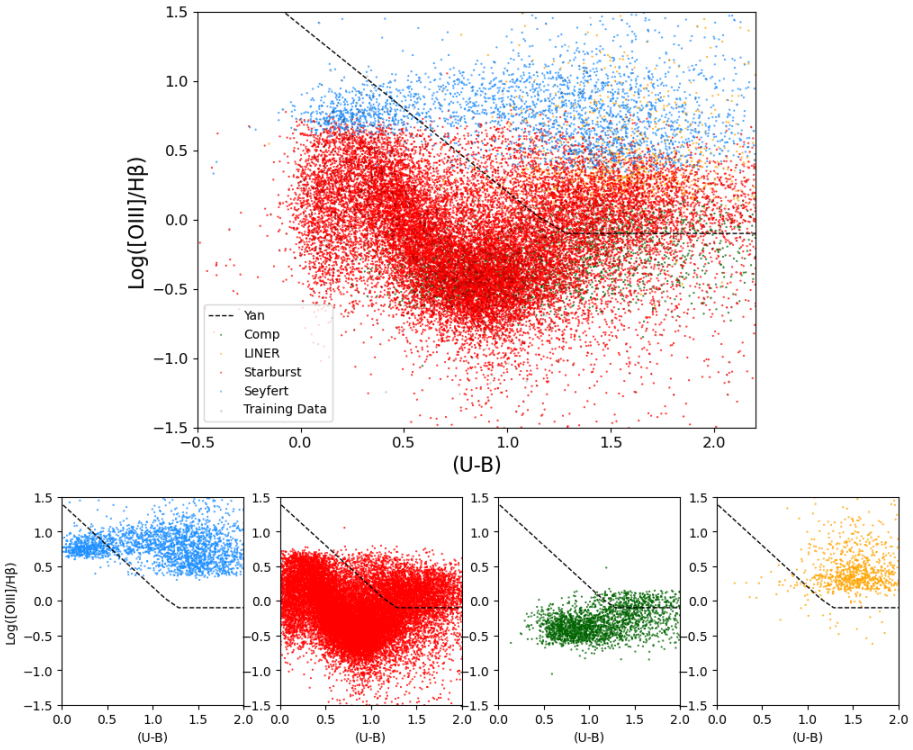


Figura 5-8: Diagrama Diagnóstico U-B generado a través de Machine Learning, representado en Python.

5.3.2 Proyección 3D

Una vez reflejados los datos en los diferentes diagramas diagnóstico en dos dimensiones, resultado del aprendizaje de las redes neuronales, se propuso realizar una proyección tridimensional, con el fin de visualizar y tener una mejor comprensión del comportamiento de las galaxias utilizando diferentes ángulos de observación.

Diagrama BPT vs. Diagrama WHAN

La **Figura 5-9** es una primera representación tridimensional, en el que el diagrama diagnóstico BPT se grafica en los ejes X y Z, mientras que el diagrama WHAN se proyecta en los ejes X y Y, en esta imagen se muestra la distribución de las galaxias a lo largo de las diferentes dimensiones, reflejando la zona en la que se agrupa cada uno de los objetos.

En la **Figura 5-10** se observa que las poblaciones de AGN se concentran en el cuadrante positivo de los ejes XZ y en el cuadrante negativo del eje Y. En contraste, las galaxias de formación estelar se dispersan hacia el eje Y positivo y negativamente en los otros dos ejes. Las galaxias compuestas y LINER se visualizan de manera difusa entre los otros dos grandes grupos, pero están situadas dentro de los límites de las curvas teóricas.

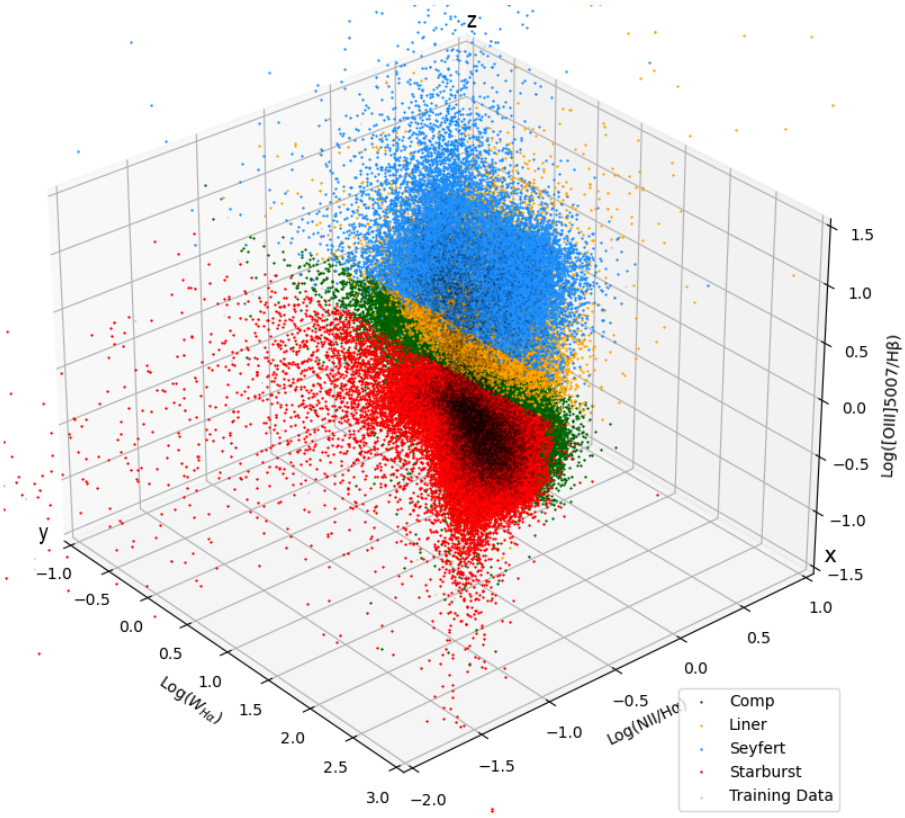


Figura 5-9: Diagrama Diagnóstico BPT vs. Diagrama Diagnóstico WHAN.

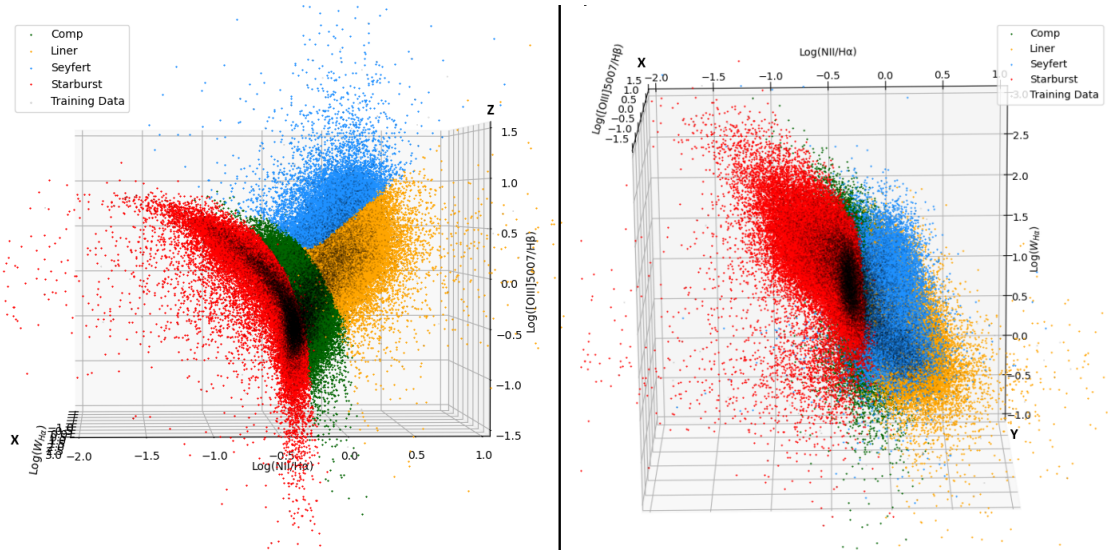


Figura 5-10: Proyecciones tridimensionales. Panel izquierdo: Diagrama Diagnóstico BPT. Panel derecho: Diagrama Diagnóstico WHAN.

Esta proyección tridimensional filtra y representa gráficamente los objetos que poseen un valor redshift no mayor al del Universo local, aproximadamente z <0.1. Se eliminaron las galaxias retiradas que se encontraban dentro del diagrama WHAN, lo que facilita la obtención de información clasificada de manera precisa y minimiza el error inherente al diagrama tradicional.

Diagrama Azul vs. Diagrama U-B

La siguiente combinación de observables se representa en la **Figura 5-11**, donde se reflejan características fotométricas del índice de color U-B en el eje YZ, mientras que el diagrama azul se grafica en la proyección de los ejes XZ. La distribución poblacional agrupa la totalidad de la muestra en estas tres dimensiones, proyectando la ubicación de cada conjunto etiquetado.

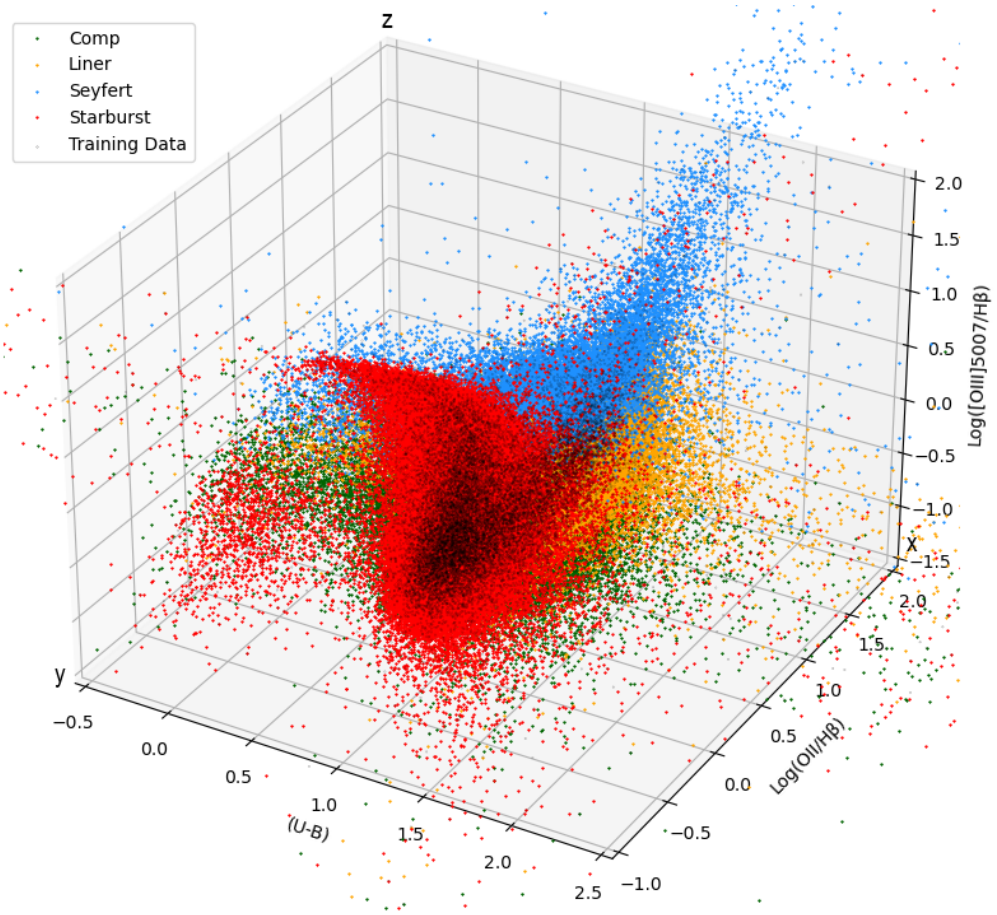


Figura 5-11: Diagrama Diagnóstico U-B vs. Diagrama Diagnóstico Azul.

Estos diagramas se centran en clasificar aquellas galaxias que poseen un redshift más allá del Universo local, es decir $z > 0.1$. Como bien se evidenció en los diagramas de dos

dimensiones, los objetos que poseen estos valores de corrimiento al rojo se agrupan siguiendo estas proyecciones, lo que conduce a la reclasificación de las galaxias inicialmente etiquetadas según el diagrama BPT.

En las proyecciones de la **Figura 5-12** se visualiza desde diferentes ángulos la distribución poblacional de cada grupo de galaxias con su correcta clasificación respecto a los diagramas U-B y Azul.

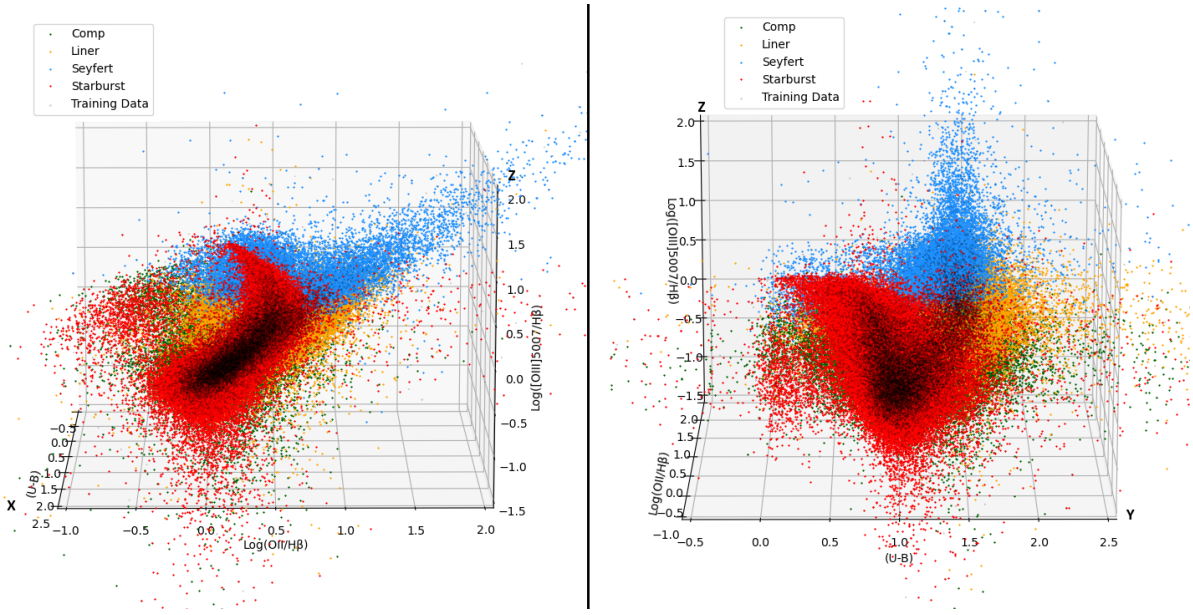


Figura 5-12: Proyecciones tridimensionales. Panel izquierdo: Diagrama Diagnóstico Azul. Panel derecho: Diagrama Diagnóstico U-B.

Diagrama Color-Color

La distribución en el espacio de parámetros tridimensional combina la fotometría de dos colores con la relación de líneas [OIII]/Hβ. En la **Figura 5-13**, se ilustra cómo los AGN de Tipo 1 se destacan de manera más efectiva entre las fuentes emisoras lineales mediante criterios fotométricos, mientras que las fuentes de tipo Starburst se pueden distinguir en función de sus propiedades espectrales.

Las galaxias Seyfert 1 se ubican en el área positiva de los filtros fotométricos u-r y g-z, como se observa en la **Figura 5-14**, esto debido a que la fuente central tiene mayor actividad hacia el espectro azul, de manera contraria los objetos Starburst se grafican en valores cercanos al cero indicando mayor actividad hacia el espectro rojo. El eje x donde se visualiza la relación Log[OIII]/Hβ realiza la separación de los dos grandes grupos, lo que permite confirmar que los filtros fotométricos son buenos trazadores para definir el tipo de actividad galáctica de la muestra poblacional.

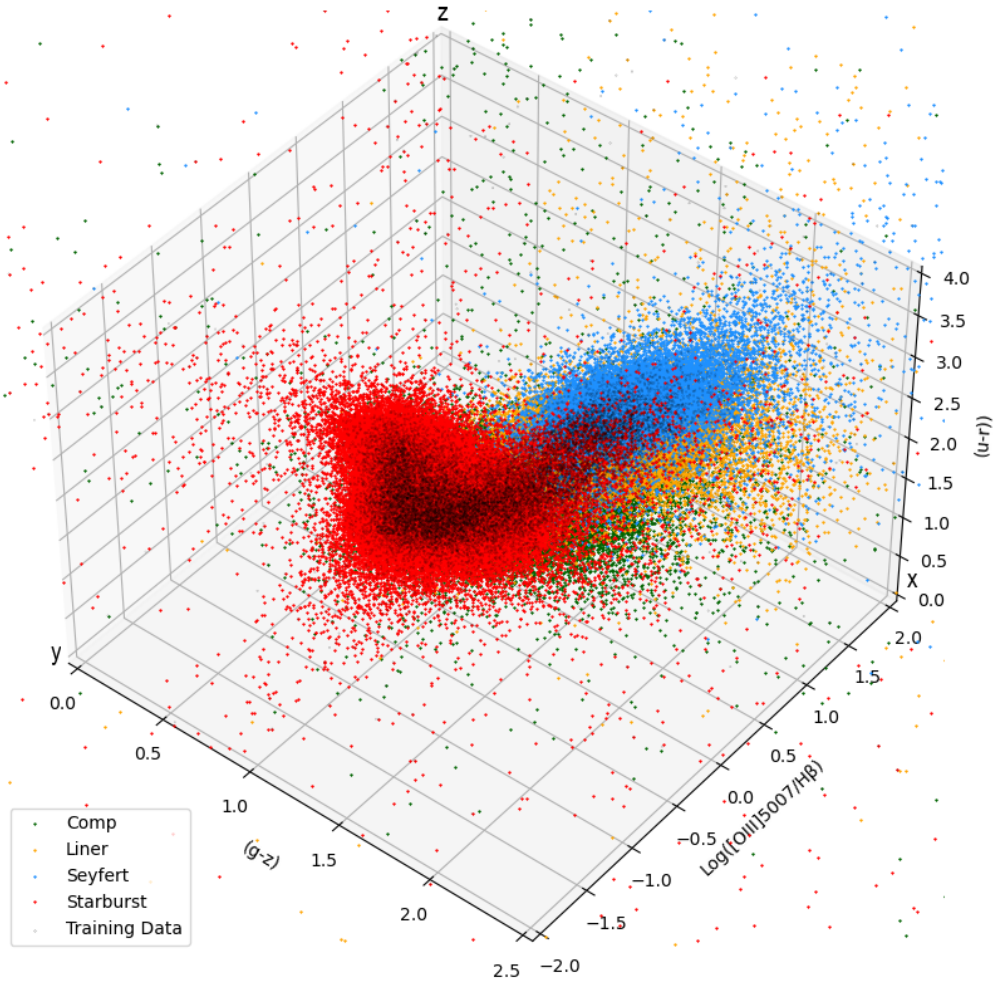


Figura 5-13: Diagrama Diagnóstico Color-Color vs. log[OIII]/Hβ.

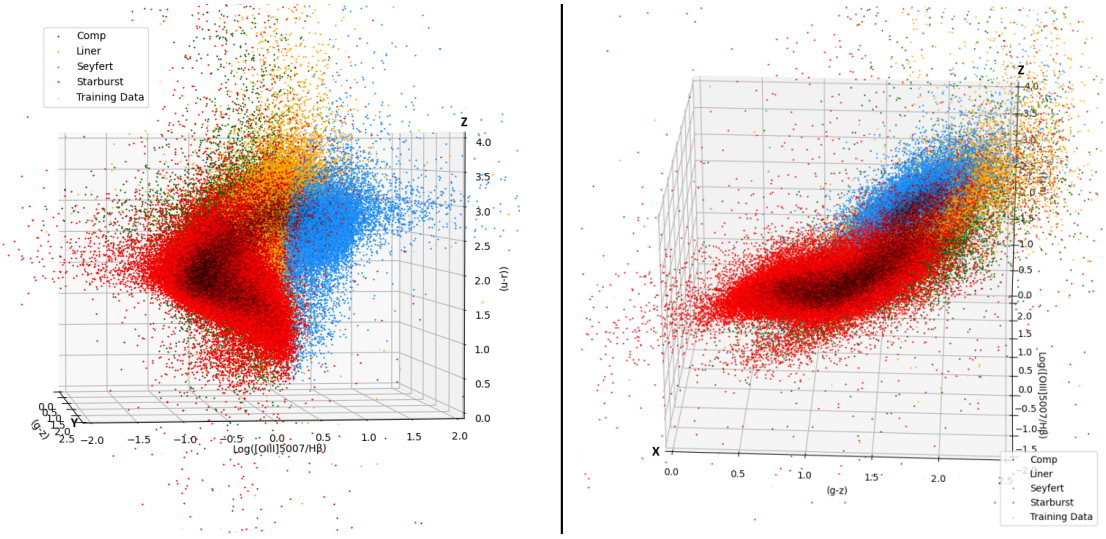


Figura 5-14: Proyecciones tridimensionales, vista desde diferentes ángulos para el Diagrama Diagnóstico Color-Color.

Es fundamental entender que cada uno de estos diagramas diagnósticos representa el conjunto total de galaxias abarcando todos los rangos de corrimiento al rojo. Por esta razón, muchas galaxias pueden cruzar los límites teóricos en los gráficos, sin embargo, aquellas que se encuentran fuera del Universo local son categorizadas según la clasificación proporcionada por el diagrama azul y el diagrama U-B.

5.4 Tabla de Resultados

En el Apéndice 3 (Tabla C.1) se tiene una visualización general de la tabla final de clasificación, obtenida como resultado del algoritmo desarrollado. Se compone de una columna inicial llamada ‘‘ID’’, la cual contiene la identificación de cada una de las galaxias. Las siguientes diez columnas representan las características de entrada utilizadas dentro la red neuronal, y, finalmente, la última columna denominada ‘‘Predicted Class’’, da como resultado la etiqueta de clasificación para cada uno de los objetos.

Esta información está disponible dentro del repositorio publicado [1].

[1] Ver repositorio Github: `https://github.com/katthe/Machine-Learning-Classification.git`

6 Conclusiones

En el transcurso de esta investigación, se examinaron detalladamente diversos métodos de aprendizaje empleados en Machine Learning. El primer método evaluado, el modelo de árbol de decisión, exhibió una tasa de efectividad del 97 %. A pesar de que los resultados ofrecen una clasificación aceptable al elegir el número adecuado de capas de profundidad, este algoritmo requiere tomar una muestra del 60 % del total de la información extraída para llevar a cabo el entrenamiento. Esta condición representa una desventaja significativa, ya que la población de galaxias destinada a la fase de prueba o testeo se reduce a tan solo el 40 %, y el objetivo principal es lograr clasificar la mayor cantidad posible de objetos de la base de datos, lo que subraya la limitación de este enfoque.

Por otra parte, se evaluó el método de aprendizaje no supervisado utilizando los datos extraídos de la base de datos, sin embargo, esta metodología no produjo los resultados esperados, ya que, al carecer de información que oriente a los centroides sobre su ubicación con respecto a las zonas de clasificación, el algoritmo selecciona una región aleatoria que no cumple con los criterios del diagrama BPT. Como resultado, se observa simplemente la agrupación de cuatro conjuntos de galaxias, pero sin una correspondencia clara entre sus características.

El aprendizaje profundo, también conocido como Deep Learning, destacó con resultados superiores en comparación con otros métodos de aprendizaje automático, razón por la cual, el algoritmo de clasificación se diseñó fundamentándose en este modelo de redes neuronales.

En la investigación, se exploraron y analizaron diversos diagramas diagnósticos propuestos por distintos autores a lo largo del tiempo. Estos diagramas, como el diagrama azul, WHAN, U-B y Color-Color, aprovechan características específicas de las galaxias activas. Su utilidad radica en la capacidad para reclasificar y depurar objetos que podrían tener una etiqueta o clasificación inicial errónea proveniente de la base de datos del SDSS, o del tradicional diagrama BPT, que utiliza líneas y curvas divisorias como fronteras entre poblaciones.

La información fundamental sobre el comportamiento de la muestra de galaxias se extrae de datos fotométricos y líneas espectrales de emisión, abarcando un amplio rango de redshift que oscila entre $0 < z < 1$. La variación en los datos recibidos se relaciona con la distancia a la que se encuentra la población de objetos, para galaxias dentro del universo local, las líneas

cuya longitud de onda se ubican hacia el rojo del espectro electromagnético demuestran un mejor comportamiento bajo el diagrama BPT y el diagrama WHAN, en contraste con las líneas situadas hacia el espectro azul y los datos fotométricos correspondiente a galaxias con alto corrimiento al rojo, que logran clasificarse con los diagramas Color-Color y el diagrama U-B.

El desarrollo de proyecciones tridimensionales de los diferentes diagramas diagnósticos fue esencial para permitir una comprensión y visualización mejoradas del comportamiento de las galaxias activas, basadas en sus características espectroscópicas y fotométricas, graficando a lo largo de los tres ejes la distribución poblacional de los objetos, permitiendo visibilizar el tipo de actividad galáctica predominante.

Los datos de entrada en una red neuronal son las variables principales que permiten que el algoritmo de aprendizaje obtenga una mayor certeza en la clasificación de resultados. Para alcanzar este objetivo, es esencial construir una red neuronal con múltiples capas ocultas, capaces de procesar una mayor cantidad de entradas y datos. La elección de la muestra de entrenamiento, la configuración de los hiperparámetros y la arquitectura de la red neuronal son la pieza clave para el funcionamiento eficaz del algoritmo. Estos aspectos permiten optimizar el modelo hasta alcanzar una precisión del 99.34 %, incluso con una muestra de entrenamiento reducida, obteniendo como resultado final un proceso que determina de manera satisfactoria la clasificación de las galaxias activas.

En futuras investigaciones, se busca continuar en la implementación de este algoritmo de clasificación, extendiéndolo a nuevas bases de datos y a la implementación de nuevos diagramas diagnóstico como el Mass-Excitation (MEx), Color-Excitation (CEx), diagrama DEW, diagrama TBT, y demás diagramas propuestos. Este avance se llevará a cabo mediante la creación e integración de una plataforma interactiva que permitirá a los usuarios elegir parámetros específicos, determinar el número de galaxias a analizar y seleccionar la información que desean clasificar, entre otras características. El objetivo es hacer que el modelo sea accesible y adaptable para cualquier persona interesada en emprender trabajos e investigaciones en esta área.

A Apéndice 1

A.1 Tabla de datos del Sloan Digital Sky Survey

specobjid	ra	dec	h_alpha	h_beta	sii	nii	oi	oii	oiii	h_alpha_eqw	u	g	r	i	z	redshift
299489676975171584	146.71421	-1.0413043	462.12	98.79	105.45	205.90	16.14	34.06	72.29	-5.45	17.15	15.50	14.68	14.23	13.90	0.02
299490502078654464	146.62857	-0.76513683	99.40	25.64	35.33	51.59	8.49	27.10	21.44	-5.13	19.50	17.59	16.68	16.21	15.85	0.06
299491051834468352	146.63167	-0.98827781	127.31	38.33	30.27	25.79	6.27	47.95	35.47	-21.19	19.48	18.33	17.84	17.53	17.42	0.05
299491326712375296	146.91945	-0.99049175	13.32	7.02	6.76	13.53	4.22	7.51	8.56	-1.84	20.44	18.65	17.47	17.00	16.67	0.21
299492700632147968	146.59272	-0.76025604	211.50	53.18	41.64	76.35	7.21	27.30	15.60	-14.73	19.07	17.62	16.97	16.58	16.35	0.07
299492975979816960	146.9635	-0.75935173	21.56	3.30	12.78	15.84	3.17	14.52	17.70	-1.43	20.32	18.51	17.62	17.21	16.91	0.10
299493800613537792	146.85983	-0.80890165	11.44	4.32	10.43	19.36	0.59	0.39	11.26	-1.19	20.78	18.65	17.31	16.71	16.22	0.13
299494350369351680	146.8577	-0.66285236	29.85	10.79	0.62	10.10	3.17	15.98	0.24	-8.90	19.90	18.33	17.61	17.21	17.15	0.08
299495724758886400	147.02336	-0.16009352	37.99	3.40	5.21	14.31	1.64	5.25	4.28	-5.10	20.49	18.88	17.79	17.29	16.88	0.22
299496274514700288	146.95607	-0.34230044	77.05	22.93	36.78	58.87	12.49	30.49	20.21	-4.77	21.02	18.92	17.61	17.03	16.59	0.13
299496823800752128	146.92201	-0.30646208	58.96	8.35	4.66	30.14	0.94	16.76	3.66	-5.15	20.13	18.56	17.49	16.99	16.62	0.13
299498473068193792	147.32951	0.028902695	493.31	108.99	86.08	314.71	21.16	81.83	56.13	-15.07	17.98	16.31	15.53	15.11	14.78	0.05
299499022824007680	147.24805	-0.035723556	20.75	7.65	18.34	21.75	0.74	4.62	12.00	-1.08	20.39	18.09	17.02	16.50	16.12	0.08
299499297701914624	147.18679	-0.49381256	74.27	11.48	21.24	50.73	15.78	7.37	18.21	-4.56	19.60	17.70	16.67	16.11	15.69	0.09
299500672561211392	146.57134	-0.95721148	351.30	66.98	48.69	147.91	7.27	20.08	18.08	-22.02	18.54	17.05	16.26	15.80	15.45	0.07
299500947439118336	146.56561	-1.0847419	169.81	24.94	98.01	248.93	37.84	39.10	135.90	-7.64	20.11	24.89	19.17	23.77	21.90	0.10
299501496725170176	146.57273	-1.0608357	129.62	28.92	20.66	49.31	7.49	19.14	5.02	-7.72	18.78	24.40	17.82	23.29	24.38	0.06
299501497194932224	146.30548	-1.1305512	332.55	73.13	51.57	94.03	25.81	66.09	26.21	-44.65	21.35	20.04	22.93	20.53	21.00	0.15
299503421340280832	146.51283	-0.84576492	1151.21	257.94	194.40	563.46	32.06	171.38	95.85	-38.93	17.37	16.18	15.50	15.06	14.79	0.06
299504245974001664	146.44831	-0.71339967	919.36	194.09	140.53	384.86	30.06	114.93	203.21	-59.88	19.02	17.93	17.35	16.89	16.70	0.11
299506169649588224	146.90232	-0.31366176	32.89	4.49	16.24	6.37	3.39	5.19	4.46	-4.20	20.13	18.09	17.01	16.49	16.11	0.13
299506170119350272	146.85757	-0.21877608	18.95	6.67	14.03	20.55	4.79	23.43	5.86	-1.40	20.76	18.75	17.56	17.02	16.62	0.12
299506444527495168	146.86432	-0.46407191	451.13	104.85	64.20	151.64	6.88	39.98	15.13	-27.88	19.18	17.88	17.19	16.73	16.47	0.07
299506449957257216	146.90125	-0.4131573	131.26	34.24	42.04	64.89	13.77	44.77	46.41	-4.76	19.95	18.12	17.12	16.62	16.22	0.06
299506719405402112	146.78847	-0.31065752	73.03	13.08	16.53	14.28	4.53	16.37	17.06	-22.13	20.66	18.72	18.10	17.77	17.64	0.06
299506994283309056	146.77711	-0.24198689	53.77	18.31	10.60	21.45	2.77	6.61	9.87	-6.95	21.05	19.10	18.07	17.59	17.24	0.13
299507269630978048	146.86816	-0.48679252	59.32	10.32	20.57	26.78	5.71	10.08	4.65	-12.31	20.36	18.97	17.94	17.38	17.04	0.08
299507818917029888	146.89158	-0.51266153	138.52	15.70	35.54	70.23	9.52	20.49	6.05	-11.65	21.20	18.96	17.88	17.29	16.86	0.08
299508643550750720	146.8833	-0.49469887	399.29	65.30	25.75	207.71	15.05	39.83	33.38	-14.43	18.61	16.86	15.80	15.23	14.82	0.08
299508918898419712	146.85967	-0.098938627	234.37	44.53	39.62	158.44	9.66	22.52	49.31	-21.21	20.37	18.89	17.96	17.44	17.07	0.13
299509193306564608	146.85676	-0.27404021	146.68	39.24	29.41	13.94	4.08	91.99	98.38	-36.21	19.68	18.52	18.01	17.75	17.52	0.02
299509468654233600	146.76078	0.015978141	123.23	18.49	25.15	52.61	2.98	11.09	11.96	-9.24	21.10	18.51	17.39	16.76	16.42	0.12
299511179216675264	146.81199	-0.19004646	16.28	7.33	3.46	15.05	5.20	4.14	25.05	-0.61	19.61	17.56	16.56	16.09	15.71	0.06
299511667207727104	146.23485	-1.1527694	615.69	115.55	117.16	186.26	19.11	91.13	47.01	-56.86	22.89	20.76	23.73	24.36	18.72	0.07
299511667677489152	146.11835	-0.8681462	204.30	39.90	37.21	107.44	6.82	29.25	19.36	-21.22	20.26	18.74	17.91	17.45	17.18	0.13
299512766719354880	146.24259	-0.76261796	477.04	141.89	30.17	41.63	6.56	157.49	549.36	-287.61	20.49	20.16	19.70	20.08	19.67	0.30
299513042067023872	146.12796	-1.1897714	38.37	10.34	11.93	18.23	2.04	10.27	4.38	-3.83	20.25	18.45	17.48	17.02	16.68	0.14
299514690864703488	146.09369	-0.79308773	31.84	5.09	25.85	50.65	3.17	26.97	16.01	-0.81	18.66	16.77	15.84	15.41	15.09	0.07
299517714521679872	146.74174	-0.52469154	167.03	40.28	12.84	128.42	7.62	17.92	56.40	-23.10	19.92	18.59	17.78	17.30	17.06	0.13
299518264277493760	146.7517	-0.40839151	1425.26	355.56	274.37	382.22	46.15	552.16	242.79	-49.33	17.81	16.61	16.01	15.66	15.39	0.04
299518539155400704	146.62966	-0.46568888	139.39	37.10	67.13	111.40	28.84	54.81	37.07	-8.00	19.38	17.63	16.71	16.24	15.88	0.09
299518814033307648	146.49052	-0.36931044	548.39	109.26	104.75	249.00	21.16	73.03	28.72	-23.64	19.20	17.67	16.75	16.27	15.94	0.12
299519914014697472	146.60365	-0.19149282	290.14	65.58	32.11	96.71	2.34	19.09	11.18	-21.08	18.92	17.66	16.98	16.56	16.32	0.12
299520188892604416	146.63068	0.070984066	33.48	4.40	17.29	32.35	7.73	11.56	21.31	-2.73	20.99	18.94	17.64	17.07	16.61	0.13
299520463770511360	146.60806	0.047679187	120.09	32.91	24.63	46.93	6.19	33.17	9.53	-16.86	20.33	18.84	18.04	17.61	17.32	0.09
299520738648418304	146.60631	-0.21322083	140.51	28.99	22.31	51.51	1.26	13.63	2.68	-16.12	20.41	19.03	18.13	17.65	17.30	0.12
299521287934470144	146.59449	-0.13317431	80.99	16.28	20.53	13.20	3.35	11.17	12.37	-18.03	19.99	18.39	17.67	17.27	16.94	0.05
299522937201911808	145.89059	-1.0976161	350.92	86.22	62.77	131.93	13.59	102.17	32.62	-19.72	17.96	16.71	16.19	15.86	15.65	0.06
299525411572836352	145.93788	-0.73396501	94.78	18.60	14.59	40.68	2.52	10.83	7.14	-8.22	19.37	17.70	16.83	16.39	16.06	0.14
299526510614702080	145.82165	-0.84656746	39.17	9.23	7.60	17.72	1.63	2.54	0.27	-4.43	20.06	18.35	17.55	17.15	16.87	0.07
299526785492609024	145.87447	-0.60874859	99.42	27.63	9.53	28.47	1.64	10.53	4.34	-6.61	18.20	16.70	16.01	15.64	15.39	0.07
299527060370515968	145.89232	-1.0226363	44.52	11.27	18.18	53.54	12.26	15.75	25.16	-2.39	20.39	18.22	17.01	16.51	16.12	0.15
299528159882143744	146.29988	-0.12001413	299.84	67.28	78.53	262.56	25.00	39.09	589.12	-7.86	17.87	15.94	14.98	14.50	14.08	0.03
299528160351905792	146.37378	-0.36838682	737.11	149.51	267.29	407.28	92.59	112.53	95.27	-30.87	19.39	17.22	16.26	15.72	15.37	0.05
299528435229812736	146.38135	-0.38646442	97.17	12.95	26.07	52.93	5.62	0.37	9.00	-8.02	20.80	18.27	17.23	16.67	16.23	0.05
299528984515864576	146.32316	-0.029989881	2296.63	615.19	156.66	80.72	30.35	198.70	2842.66	-366.37	18.07	17.12	16.74	16.60	16.48	0.02
299529259393771520	146.36979	0.082264176	104.72	27.18	26.52	41.91	4.42	16.89	7.45	-12.26	19.77	18.23	17.46	16.97	20.41	0.07
299529534741440512	146.35204	-0.33269118	1334.97	266.45	210.07	489.74	44.57	184.16	86.14	-36.25	18.74	17.25	16.49	16.02	15.72	0.05

B Apéndice 2

B.1 Tabla de Características

OIII/Hβ	NII/Hα	SII/Hα	OI/Hα	OII/Hα	OII/Hβ	U-B	u-r	g-z	W$_H\alpha$
-0.077693	-0.160493	-0.241588	-1.068479	-0.176791	0.411677	1.467085	2.82	1.74	0.710117
-0.033678	-0.569062	-0.356752	-1.307595	-0.119436	0.401888	0.871169	1.64	0.91	1.326131
0.086137	0.131009	-0.090277	-0.499192	-0.000653	0.277515	1.372993	2.97	1.98	0.264818
-0.532624	-0.318160	-0.479468	-1.467375	-0.551378	0.048184	1.106399	2.10	1.27	1.168203
0.729459	-0.009572	-0.035888	-0.832589	-0.102434	0.712701	1.388675	2.70	1.60	0.155336
0.416055	0.352857	0.162466	-1.287574	0.195638	0.618581	1.639587	3.47	2.43	0.075547
-1.652810	-0.346222	-0.660031	-0.973885	-0.102216	0.339706	1.200491	2.29	1.18	0.949390
0.099965	-0.299546	-0.796481	-1.364825	-0.679302	0.368888	1.231855	2.70	2.00	0.707570
-0.054838	0.007488	-0.018187	-0.790210	-0.036126	0.490242	1.616064	3.41	2.33	0.678518
-0.358205	-0.167088	-0.598247	-1.797430	-0.407510	0.441361	1.200491	2.64	1.94	0.711807
-0.288192	-0.070854	-0.511019	-1.367604	-0.408983	0.246751	1.278901	2.45	1.53	1.178113
0.195520	0.144780	0.151181	-1.447786	-0.270523	0.162834	1.772884	3.37	1.97	0.033424
0.200368	-0.041188	-0.315719	-0.672706	-0.639089	0.171782	1.459244	2.93	2.01	0.658965
-0.568747	-0.251322	-0.613001	-1.684144	-0.760704	-0.040971	1.137763	2.28	1.60	1.342817
0.736323	0.290468	0.012156	-0.652012	-0.039660	0.793407	-3.778544	0.94	2.99	0.883093
-0.760495	-0.295371	-0.533460	-1.238190	-0.675397	-0.023923	-4.437188	0.96	0.02	0.887617
-0.445629	-0.424253	-0.605719	-1.110069	-0.454861	0.202901	0.996625	-1.58	-0.96	1.649821
-0.429927	-0.185937	-0.519475	-1.555191	-0.565291	0.084345	0.902533	1.87	1.39	1.590284
0.019942	-0.253826	-0.563604	-1.485497	-0.605627	0.069855	0.824123	1.67	1.23	1.777282
-0.135635	-0.226735	-0.345628	-1.456851	-0.131924	0.538118	1.263219	2.47	1.60	0.736397
-0.002911	-0.588668	-0.264211	-0.986864	-0.334364	0.530454	1.569018	3.12	1.98	0.623249
-0.056228	0.159507	0.010416	-0.597274	0.177236	0.630719	1.545495	3.20	2.13	0.146128
-0.840730	-0.349144	-0.627809	-1.816713	-0.715682	-0.081949	0.988784	1.99	1.41	1.445293
0.132078	-0.181619	-0.269389	-0.979198	-0.144312	0.439287	1.404357	2.83	1.90	0.677607
0.115371	-0.584519	-0.387975	-1.207403	-0.220641	0.526252	1.490608	2.56	1.08	1.344981
-0.268371	-0.274782	-0.417728	-1.288060	-0.670977	-0.203125	1.498449	2.98	1.86	0.841985
-0.346227	-0.221020	-0.277241	-1.016565	-0.308414	0.451108	1.059353	2.42	1.93	1.090258
-0.414144	-0.170654	-0.391776	-1.162876	-0.668610	0.277003	1.725838	3.32	2.10	1.066326
-0.291427	-0.159468	-0.810863	-1.423752	-0.876030	-0.089655	1.341629	2.81	2.04	1.159266
0.044282	-0.045681	-0.517178	-1.384925	-0.817477	-0.096228	1.129922	2.41	1.82	1.326541

C Apéndice 3

C.1 Tabla Final de Clasificación

ID	OIII/Hβ	NII/Hα	SII/Hα	OI/Hα	OII/Hα	OII/Hβ	U-B	u-r	g-z	W$_{H\alpha}$	Predicted Class
299492700632147968	-0.53267	-0.318124	-0.479489	-1.46731	-0.551391	0.0481561	1.11076	2.09694	1.26733	1.16826	Starburst
299502871114704896	-0.599074	-0.246748	-0.578291	-1.65287	-0.699119	0.000874148	1.13917	2.28749	1.60053	1.33257	Comp
299579836928649216	-0.362337	-0.373545	-0.48585	-1.40116	-0.423432	0.170684	0.923508	1.81493	1.09431	1.56762	Starburst
299661201258866688	-0.00432159	-0.484041	-0.387146	-1.28653	-0.0737924	0.552053	0.716827	1.31492	0.7326	1.53358	Starburst
299663675160029184	-0.513466	-0.211681	-0.587289	-1.91448	-0.728978	-0.108602	1.12216	2.20992	1.44853	0.670418	Comp
300633444046628864	-0.746334	-0.25587	-0.547139	-1.75177	-0.788431	-0.141102	0.962909	1.87961	1.23879	1.39259	Starburst
300685945726855168	-0.558125	-0.404803	-0.473479	-1.51421	-0.441456	0.189224	0.891587	1.66396	0.99217	1.5703	Starburst
300700239378016256	-0.408642	-0.361765	-0.391423	-1.22038	-0.486056	0.152456	1.16397	2.27093	1.49873	1.14372	Starburst
300760987395450880	-0.632709	-0.315916	-0.645059	-1.44639	-0.746022	-0.145728	1.00573	2.08823	1.46921	1.04024	Starburst
300776105680332800	-0.514647	-0.211487	-0.560573	-1.51333	-0.440775	0.150944	1.04132	2.03654	1.28	0.911743	Comp
301784083384526848	-0.267357	-0.181223	-0.55186	-1.56033	-0.561281	-0.0294325	1.27319	2.53981	1.63651	0.578161	Comp
301889911378700288	0.140244	0.262108	-0.0322834	-0.622469	-0.0375011	0.560967	1.43476	2.70538	1.59175	0.629384	LINER
302933902316562432	-0.35008	-0.193469	-0.353924	-1.30512	-0.136845	0.404174	1.1676	2.23782	1.40226	1.00562	Comp
302949020601444352	-0.453723	-0.221908	-0.482399	-1.08777	-0.403223	0.329645	1.27753	2.5318	1.66866	1.13385	Comp
302950120113072128	-0.315931	-0.529579	-0.328228	-1.469	-0.00652824	0.472964	0.849136	1.47993	0.72044	1.20765	Starburst
303002621793298432	-0.0947901	-0.472977	-0.45865	-1.43699	-0.0789645	0.451674	0.715353	1.31103	0.73899	1.70856	Starburst
303006470083995648	-0.655331	-0.326704	-0.509028	-1.79878	-0.56063	0.0466545	0.902682	1.79481	1.13172	1.16153	Starburst
303017740078180352	-0.280861	-0.190651	-0.436854	-1.91603	-0.424589	0.337515	1.37687	2.72271	1.80858	0.951131	Comp
303021588368877568	0.219769	0.234658	-0.0693658	-0.470091	0.146816	0.782828	1.48592	2.83372	1.66571	0.25485	LINER
304079593415927808	-1.13976	-0.331539	-0.622764	-1.5629	-0.64369	0.0283153	0.988596	1.9708	1.35354	1.16232	Starburst
305183227776100352	-0.340791	-0.697341	-0.160436	-1.06525	0.128652	0.488121	0.921634	1.64001	0.6045	0.952377	Starburst
305186801188890624	0.198682	0.115767	0.103099	-0.981955	0.180375	0.697482	1.48421	2.91222	1.75784	0.0497822	LINER
305214288979585024	0.186354	0.0438142	0.0238402	-0.785634	0.103956	0.610412	1.60281	3.1425	1.9421	0.207742	LINER
305224734340048896	-0.348173	-0.269076	-0.413052	-1.56181	-0.424095	0.173775	1.15089	2.27024	1.52161	1.15766	Comp
305280809433065472	-0.161768	-0.0742301	-0.308531	-1.25356	-0.187961	0.25949	1.40969	2.73881	1.62942	0.422179	Comp
306326995719972864	-0.332329	-0.38788	-0.526844	-1.52849	-0.2295	0.327872	0.847247	1.57445	0.95652	1.57521	Starburst
306330019376949248	-0.198796	-0.275349	-0.333721	-1.24353	-0.329437	0.458268	1.32947	2.62611	1.79469	1.0433	Comp
306336891324622848	-0.120577	-0.441667	-0.433427	-1.50461	-0.500712	0.327758	1.37048	2.52846	1.47765	1.25641	Starburst
306338815469971456	-0.707175	-0.345772	-0.52316	-1.52595	-0.70975	0.0254668	1.12549	2.29168	1.64857	1.38751	Starburst

Referencias Bibliográficas

Alpaydin, E.: , 2014; *Introduction to Machine Learning*; Adaptive Computation and Machine Learning; MIT Press, Cambridge, MA; 3ª edición; ISBN 978-0-262-02818-9.

Andrew, F.: , 1999; Active galactic nuclei; en *Proceedings of the National Academy of Sciences of the United States of America*, tomo 96; págs. 4749--51.

Antonucci, R.: , 1993; Unified models for active galactic nuclei and quasars.; ; **31**: 473--521; doi:10.1146/annurev.aa.31.090193.002353.

Baldwin, J. A.; **Phillips, M. M.** & **Terlevich, R.**: , 1981; Classification parameters for the emission-line spectra of extragalactic objects.; ; **93**: 5--19; doi:10.1086/130766.

Bishop, C.: , 2013; *Pattern Recognition and Machine Learning: All "just the Facts 101"Material*; Information science and statistics; Springer (India) Private Limited; ISBN 9788132209065; URL `https://books.google.com.co/books?id=HL4HrgEACAAJ`.

Blanton, M. R. & **Roweis, S.**: , 2007; K-corrections and filter transformations in the ultraviolet, optical, and near-infrared; *The Astronomical Journal*; **133** (2): 734–754; doi:10.1086/510127; URL `http://dx.doi.org/10.1086/510127`.

Brinchmann, J.; **Charlot, S.**; **White, S. D. M.**; **Tremonti, C.**; **Kauffmann, G.**; **Heckman, T.** & **Brinkmann, J.**: , 2004; The physical properties of star-forming galaxies in the low-redshift Universe; ; **351** (4): 1151--1179; doi:10.1111/j.1365-2966.2004.07881.x.

Centeno Franco, A.: , 2019; *DEEP LEARNING*; https://hdl.handle.net/11441/90004; Universidad de Sevilla.

Chagas, E. T. D. O.: , 2019; Aprendizaje profundo y sus aplicaciones hoy; *Revista Científica Multidisciplinar Núcleo do Conhecimento*; **4** (5): 05--26; doi:ISSN:2448-0959.

Cid Fernandes, R.; **Stasinska, G.**; **Asari, N. V.**; **Mateus, A.**; **Schlickmann, M.** & **Schoenell, W.**: , 2010a; Emission line taxonomy and the nature of AGN-looking galaxies in the SDSS; *IAU Symp.*; **267**: 65; doi:10.1017/S1743921310005582.

Cid Fernandes, R.; **Stasińska, G.**; **Schlickmann, M. S.**; **Mateus, A.**; **Vale Asari, N.**; **Schoenell, W.** & **Sodré, L.**: , 2010b; Alternative diagnostic diagrams and the

'forgotten' population of weak line galaxies in the sdss: The forgotten population of wlgs in the sdss; *Monthly Notices of the Royal Astronomical Society*; **403** (2): 1036–1053; doi:10.1111/j.1365-2966.2009.16185.x; URL http://dx.doi.org/10.1111/j.1365-2966.2009.16185.x.

Cid Fernandes, R.; **Stasińska, G.**; **Mateus, A.** & **Vale Asari, N.**: , 2011; A comprehensive classification of galaxies in the Sloan Digital Sky Survey: how to tell true from fake AGN?; ; **413** (3): 1687--1699; doi:10.1111/j.1365-2966.2011.18244.x.

Cutiva Alvarez, K. A.: , 2018; *Estudio de una posible correlación entre la tasa de formación estelar y el tamaño de la BLR en AGNs*; Proyecto Fin de Carrera; Universidad Nacional de Colombia.

Duchi, J.; **Hazan, E.** & **Singer, Y.**: , 2011; Adaptive subgradient methods for online learning and stochastic optimization; *Journal of Machine Learning Research*; **12**: 2121--2159.

Glorot, X. & **Bengio, Y.**: , 2010; Understanding the difficulty of training deep feedforward neural networks; en *International Conference on Artificial Intelligence and Statistics*; URL https://api.semanticscholar.org/CorpusID:5575601.

Gordon-Rodriguez, E.; **Loaiza-Ganem, G.**; **Pleiss, G.** & **Cunningham, J. P.**: , 2020; Uses and abuses of the cross-entropy loss: Case studies in modern deep learning.

Hardcastle, M. J.; **Williams, W. L.**; **Best, P. N.**; **Croston, J. H.**; **Duncan, K. J.**; **Röttgering, H. J. A.**; **Sabater, J.**; **Shimwell, T. W.**; **Tasse, C.**; **Callingham, J. R.**; **Cochrane, R. K.**; **de Gasperin, F.**; **Gürkan, G.**; **Jarvis, M. J.**; **Mahatma, V.**; **Miley, G. K.**; **Mingo, B.**; **Mooney, S.**; **Morabito, L. K.**; **O'Sullivan, S. P.**; **Prandoni, I.**; **Shulevski, A.** & **Smith, D. J. B.**: , 2019; Radio-loud AGN in the first LoTSS data release. The lifetimes and environmental impact of jet-driven sources; ; **622**: A12; doi:10.1051/0004-6361/201833893.

He, K.; **Zhang, X.**; **Ren, S.** & **Sun, J.**: , 2015; Delving deep into rectifiers: Surpassing human-level performance on imagenet classification.

Heckman, T. M.: , 2005; Local starbursts in a cosmological context; *Starbursts*: 3–10; doi:10.1007/1-4020-3539-x_1; URL http://dx.doi.org/10.1007/1-4020-3539-X_1.

Higuera, M. A.: , 2012; *Intensa formación estelar en núcleos activos de galaxias, trazada por emisión de Hidrocarburos Aromáticos Policíclicos y análisis del toroide como región en donde toma lugar esta actividad estelar;*; Tesis Doctoral; National University of Colombia, Colombia.

Hinton, G.; **Deng, L.**; **Yu, D.**; **Dahl, G. E.**; **Mohamed, A.-r.**; **Jaitly, N.**; **Senior, A.**; **Vanhoucke, V.**; **Nguyen, P.**; **Sainath, T. N.** & **Kingsbury, B.**: , 2012; Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups; *IEEE Signal Processing Magazine*; **29** (6): 82--97; doi:10.1109/MSP.2012.2205597.

Hutter, F.; **Kotthoff, L.** & **Vanschoren, J.**, (Eds.): , 2019; *Automatic Machine Learning: Methods, Systems, Challenges*; Springer.

Jierula, A.; **Wang, S.**; **OH, T.-M.** & **Wang, P.**: , 2021; Study on accuracy metrics for evaluating the predictions of damage locations in deep piles using artificial neural networks with acoustic emission data; *Applied Sciences*; **11** (5); doi:10.3390/app11052314; URL `https://www.mdpi.com/2076-3417/11/5/2314`.

John Joseph, F. J.; **Nonsiri, S.** & **Monsakul, A.**: , 2021; *Keras and TensorFlow: A Hands-On Experience*; ISBN 978-3-030-66518-0; doi:10.1007/978-3-030-66519-7_4.

Kauffmann, G.; **Heckman, T. M.**; **Tremonti, C.**; **Brinchmann, J.**; **Charlot, S.**; **White, S. D. M.**; **Ridgway, S. E.**; **Brinkmann, J.**; **Fukugita, M.**; **Hall, P. B.**; **Ivezić, Ž.**; **Richards, G. T.** & **Schneider, D. P.**: , 2003; The host galaxies of active galactic nuclei; ; **346** (4): 1055--1077; doi:10.1111/j.1365-2966.2003.07154.x.

Kewley, L.; **Dopita, M.**; **Sutherland, R.**; **Heisler, C.** & **Trevena, J.**: , 2001; Theoretical modeling of starburst galaxies; *The Astrophysical Journal*; **556**; doi:10.1086/321545.

Kewley, L. J.; **Groves, B.**; **Kauffmann, G.** & **Heckman, T.**: , 2006; The host galaxies and classification of active galactic nuclei; ; **372** (3): 961--976; doi:10.1111/j.1365-2966.2006.10859.x.

Kingma, D. P. & **Ba, J.**: , 2017; Adam: A method for stochastic optimization.

Lamareille, F.: , 2010; Spectral classification of emission-line galaxies from the Sloan Digital Sky Survey. I. An improved classification for high-redshift galaxies; ; **509**: A53; doi:10.1051/0004-6361/200913168.

Lecun, Y.; **Bottou, L.**; **Orr, G.** & **Müller, K.-R.**: , 2000; Efficient backprop.

Lederer, J.: , 2021; Activation functions in artificial neural networks: A systematic overview; *CoRR*; **abs/2101.09957**; URL `https://arxiv.org/abs/2101.09957`.

Li, H.; **Krcek, M.** & **Perin, G.**: , 2020; *A Comparison of Weight Initializers in Deep Learning-Based Side-Channel Analysis*; ISBN 978-3-030-61637-3; doi:10.1007/978-3-030-61638-0_8.

Li, Y. & **Wu, H.**: , 2012; A clustering method based on k-means algorithm; *Physics Procedia*; **25**: 1104--1109; doi:https://doi.org/10.1016/j.phpro.2012.03.206; URL `https://www.sciencedirect.com/science/article/pii/S1875389212006220`; international Conference on Solid State Devices and Materials Science, April 1-2, 2012, Macao.

Liu, Q. & **Wu, Y.**: , 2012; Supervised learning; doi:10.1007/978-1-4419-1428-6_451.

McCulloch, W. S. & **Pitts, W.**: , 1943; A logical calculus of the ideas immanent in nervous activity; *The Bulletin of Mathematical Biophysics*; **5** (4): 115--133; doi:10.1007/bf02478259.

McLure, R. J.; **Jarvis, M. J.**; **Targett, T. A.**; **Dunlop, J. S.** & **Best, P. N.**: , 2006; On the evolution of the black hole: spheroid mass ratio; ; **368** (3): 1395--1403; doi:10.1111/j.1365-2966.2006.10228.x.

Mura, G. L.; **Berton, M.**; **Chen, S.**; **Chougule, A.**; **Ciroi, S.**; **Congiu, E.**; **Cracco, V.**; **Frezzato, M.**; **Mordini, S.** & **Rafanelli, P.**: , 2017; Models of emission line profiles and spectral energy distributions to characterize the multi-frequency properties of active galactic nuclei.

Naeem, S.; **Ali, A.**; **Anam, S.** & **Ahmed, M.**: , 2023; An unsupervised machine learning algorithms: Comprehensive review; *IJCDS Journal*; **13**: 911--921; doi:10.12785/ijcds/130172.

Nwankpa, C.; **Ijomah, W.**; **Gachagan, A.** & **Marshall, S.**: , 2020; Activation functions: Comparison of trends in practice and research for deep learning.

Olave Rojas, D. E.: , 2014; *Búsqueda de gradientes de metalicidad en las colas de marea de NGC 6845*; Proyecto Fin de Carrera; Universidad de La Serena, Chile.

Osterbrock, D. E.: , 1989; *Astrophysics of gaseous nebulae and active galactic nuclei.*

Padovani, P.; **Alexander, D. M.**; **Assef, R. J.**; **De Marco, B.**; **Giommi, P.**; **Hickox, R. C.**; **Richards, G. T.**; **Smolčić, V.**; **Hatziminaoglou, E.**; **Mainieri, V.** & **et al.**: , 2017; Active galactic nuclei: what's in a name?; *The Astronomy and Astrophysics Review*; **25** (1); doi:10.1007/s00159-017-0102-9; URL `http://dx.doi.org/10.1007/s00159-017-0102-9`.

Pedregosa, F.; **Varoquaux, G.**; **Gramfort, A.**; **Michel, V.**; **Thirion, B.**; **Grisel, O.**; **Blondel, M.**; **Müller, A.**; **Nothman, J.**; **Louppe, G.**; **Prettenhofer, P.**; **Weiss, R.**; **Dubourg, V.**; **Vanderplas, J.**; **Passos, A.**; **Cournapeau, D.**; **Brucher, M.**; **Perrot, M.** & **Édouard Duchesnay**: , 2018; Scikit-learn: Machine learning in python.

Peterson, B. M.: , 1997; *An Introduction to Active Galactic Nuclei.*

Rueda Vargas, S. C.: , 2020; Caracterización de agn y estimación de la masa de su agujero negro central usando espectroscopía en el rango óptico; URL `http://hdl.handle.net/1992/48881`.

Song YY, L. Y.: , 2015; Decision tree methods: applications for classification and prediction: 3; doi:10.11919/j.issn.1002-0829.215044.PMID:26120265;PMCID:PMC4466856.

Urry, M.: , 2003; The AGN Paradigm for Radio-Loud Objects; en *Active Galactic Nuclei: From Central Engine to Host Galaxy*, tomo 290 de *Astronomical Society of the Pacific Conference Series* (Editado por **Collin, S.**; **Combes, F.** & **Shlosman, I.**); pág. 3.

Vaona, L.; **Ciroi, S.**; **Di Mille, F.**; **Cracco, V.**; **La Mura, G.** & **Rafanelli, P.**: , 2012; Spectral properties of the narrow-line region in seyfert galaxies selected from the sdss-dr7: Spectral properties of the nlr in seyfert galaxies; *Monthly Notices of the Royal Astronomical Society*; **427** (2): 1266–1283; doi:10.1111/j.1365-2966.2012.22060.x; URL `http://dx.doi.org/10.1111/j.1365-2966.2012.22060.x`.

Veilleux, S. & **Osterbrock, D. E.**: , 1987; Spectral Classification of Emission-Line Galaxies; ; **63**: 295; doi:10.1086/191166.

Yan, R.; **Ho, L. C.**; **Newman, J. A.**; **Coil, A. L.**; **Willmer, C. N. A.**; **Laird, E. S.**; **Georgakakis, A.**; **Aird, J.**; **Barmby, P.**; **Bundy, K.**; **Cooper, M. C.**; **Davis, M.**; **Faber, S. M.**; **Fang, T.**; **Griffith, R. L.**; **Koekemoer, A. M.**; **Koo, D. C.**; **Nandra, K.**; **Park, S. Q.**; **Sarajedini, V. L.**; **Weiner, B. J.** & **Willner, S. P.**: , 2011; AEGIS: Demographics of X-ray and Optically Selected Active Galactic Nuclei; ; **728** (1): 38; doi:10.1088/0004-637X/728/1/38.

Zubair, M., I. M. D. A. S. A. C. M. J. M. M. M. A. . S. I. H.: , 2022; An improved k-means clustering algorithm towards an efficient data-driven modeling; *Annals of Data Science*: 1--20; doi:https://doi.org/10.1007/s40745-022-00428-2.