ANALYSIS OF MACHINE LEARNING ALGORITHMS APPLIED TO THE CLASSIFICATION OF CELESTIAL OBJECT: STARS/QUASARS/GALAXIES

ANÁLISIS DE ALGORITMOS DE APRENDIZAJE AUTOMÁTICO APLICADOS A LA CLASIFICACIÓN DE OBJETOS CELESTES: ESTRELLAS/QUÁSARES/GALAXIAS

Antonio Noguerón-Bárcenas[1]
Armando Arredondo-Valle[1]
Regina Alexia Blas Flores[1]
Jorge Abraham Vega-Méndez[1]
Emiliano Vivas Rodriguez[1]
Laura Arantzha Santos Flores[1]
Ricardo Rodriguez Figueroa[1]
Atoany Nazareth Fierro-Radilla[1]*

Tópico del congreso: Ingeniería computacional
Subtópico del congreso: Aprendizaje automático

Abstract

Stars, galaxies and quasars had been classified using morphological information, however, the farther is the object the more difficult is to obtain information. Thanks to the success of machine learning in finding patterns in data, the classification of these stellar objects can be done using another type of information such as the optical spectra. Thus, it is possible to analyze objects millions of light-years away from Earth and understand the origin of the Universe. In this paper we analyze four machine learning algorithms such as logistic regression, support vector machine, K-nearest neighbors and decision trees, on two catalogues, SDSS-DR16 and SDSS-DR17. These catalogues describe stellar objects using 17 features, however, we only used the most 10 relevant to classification. For training, we split the dataset into training set (70%) and validation set (30%). For evaluation, we used the most commonly used classification metrics such as Precision, Recall, F1-Score and Accuracy. The algorithms were compared to each other in order to know which one performs better the classification task of stars, galaxies and quasars.

Key Words: Stars, Quasars, Galaxy, Classification, Machine Learning

[1] Departamento de Computación, Departamento de Mecatrónica, Tecnológico de Monterrey

* *Autor corresponsal:* Departamento de Mecatrónica, Tecnológico de Monterrey. México 95D km 104, Real del Puente, Xochitepec, Morelos. 62790. México. Email: afierror@tec.mx

Resumen

Las estrellas, galaxias y quásares habían sido clasificados utilizando información morfológica, sin embargo, entre más lejano esté el objeto, más difícil es extraer información de él. Gracias al éxito del aprendizaje automático en encontrar patrones en los datos, la clasificación de éstos objetos estelares puede ser realizada utilizando otro tipo de información como por el ejemplo su propio espectro óptico. Así, es posible analizar objetos que se encuentran a millones de años luz de la Tierra y entender el origen del Universo. En este artículo analizamos cuatro algoritmos de aprendizaje automático como por ejemplo regresión logística, máquinas de soporte vectorial, los K-vecinos más cercanos y árboles de decisión, en dos catálogos, SDSS-DR16 y SDSS-DR17. Estos catálogos describen los objetos estelares utilizando 17 características, sin embargo, utilizamos las 10 más relevantes a la clasificación. Para el entrenamiento, dividimos el conjunto de datos en dos, en el de entrenamiento (70%) y en el de validación (30%). Para la evaluación, utilizamos las métricas más utilizadas en algoritmos de clasificación como, por ejemplo, Precisión, Recall, Métrica F1 y la Exactitud. Los algoritmos fueron comparados entre sí para determinar cual de todos es mejor en la clasificación de estrellas, galaxias y quásares.

Palabras clave: Estrellas, Cuásar, Galaxia, Clasificación, Aprendizaje Automático

Introducción

Comprender el origen de la vida ha resultado ser una de las mayores incógnitas de la humanidad. Desde que los primeros filósofos volvieron la cabeza hacia las estrellas, la clasificación de los cuerpos celestes ha sido meramente una interpretación visual. Actualmente, la Astronomía y la Cosmología se encargan del estudio y la caracterización de millones de objetos, como por ejemplo estrellas, galaxias, quásares (QSO), nebulosas planetarias, supernovas, entre otras (Robertson et al., 2015); los cuales se pueden identificar con su propio espectro óptico (Bai et al., 2018). Para realizar este estudio, se recaba información espectral utilizando grandes telescopios los cuales generan una colección de datos llamados catálogos, como por ejemplo el Estudio Digital del Cielo Sloan (Sloan Digital Sky Survey, en inglés) o la misión Gaia que está diseñada para realizar la clasificación de estrellas en nuestra galaxia y objetos extra galácticos (Gaia Collaboration et al., 2016). Hoy en día, la Astronomía está muy interesada en la clasificación de galaxias, estrellas y quásares debido a que aporta información de la física que gobierna estos objetos y cómo se creó nuestra galaxia y el universo (Clarke et al., 2020). Tradicionalmente, la información morfológica es utilizada para la clasificación de estrellas, galaxias y quásares, sin embargo, entre más lejano está el objeto de nuestros telescopios, más difícil es extraer información debido a la baja resolución de las imágenes (Vasconcellos et al., 2011). Otra manera de clasificar estos objetos estelares es utilizando magnitudes y criterios de color en un espacio multidimensional (Bai et al., 2018). La gran cantidad de información de objetos recabada por telescopios ha hecho que la tarea de clasificación manual sea casi imposible (Bai et al., 2018; Krakowski et al. 2016; Xiao y Jin, 2020). En la última década, el aprendizaje automático (machine learning) ha demostrado ser una excelente técnica para la clasificación de grandes bases de datos. Los algoritmos con bases en el aprendizaje automático, permiten el procesamiento de los datos y la obtención de respuestas a muchas preguntas en Astronomía. Particularmente, la clasificación de galaxias/estrellas/QSO se ha realizado utilizando varias técnicas de aprendizaje automático, como por ejemplo las Máquinas de Soporte Vectorial (SVM) (Solarz et al., 2012; Fadely et al., 2012), Árboles de Decision (DT) y Bosques Aleatorios (RF) (Breiman, 2001; Gal et al,. 2004; Ball et al., 2006; Vasconcellos et al., 2011), Redes Neuronales Artificiales (ANN) (Kim & Brunner 2016). El aprendizaje automático enseña a las computadoras a aprender de la "experiencia" sin ser explícitamente programadas. Las

computadoras son programadas para encontrar patrones dentro de los datos y generar una predicción en base a una correlación de variables (Bai et al., 2018). La contribución de este artículo incluye lo siguiente:

- Revisión de cuatro algoritmos de aprendizaje automático para la clasificación de estrellas, galaxias y quásares.
- Experimentación de los cuatro algoritmos utilizando dos catálogos de observaciones de cuerpos estelares, como por ejemplo el SDSS-DR16 y el SDSS-DR17.
- Evaluación de los algoritmos de aprendizaje automático
- Comparación de los algoritmos de aprendizaje automático

Aprendizaje Automático

La inteligencia artificial es la capacidad que tienen las computadoras de realizar tareas normalmente atribuidas a los seres humanos. Dentro de la inteligencia artificial, encontramos al aprendizaje automático, el cual es una disciplina, la cual, consiste en programar a la computadora para que aprenda con la experiencia (Breiman, 2001; Gal et al,. 2004; Ball et al., 2006). En otras palabras, estos algoritmos están programados de manera implícita para encontrar patrones dentro de los datos, calculando correlaciones entre variables. Estos algoritmos representan toda una rama de nuevas herramientas para identificar comportamientos complejos no lineales dentro de grandes conjuntos de datos (Costa et al., 2019). A continuación, se presentan los cuatro algoritmos que se analizaron en este artículo para la clasificación de cuerpos estelares.

Regresión logística

Cuando se trata de probar la relación bivariante de un modelo (es decir, la relación entre dos variables) la manera más conveniente es mediante el uso de modelos de regresión (Zou et al., 2019). A diferencia de la regresión lineal, la regresión logística se enfoca más en la clasificación. Este tipo de regresión utiliza un arreglo de puntos que no se registran de forma lineal, sino en datos dispersos. Cada acumulado de puntos representa una categoría, y cada punto está etiquetado en la categoría a la que pertenece. La regresión logística recibe un número arbitrario de números de entrada y obteniendo como salida una clasificación de estos, se puede representar con varias funciones, más comúnmente con la función sigmoide. Esta técnica estadística se centra en demostrar, cómo su nombre lo indica, la relación entre una variable predictora (independiente) de un resultado (variable dependiente) y que permite obtener resultados pronosticables, pero a su vez justificativos.

Máquinas de Soporte Vectorial (SVM)

La Máquina de Soporte Vectorial (SVM, por sus siglas en inglés), es considerada como uno de los algoritmos de aprendizaje automático más utilizado. Existen dos tipos de patrones. Lineales y no lineales; los patrones lineales son fáciles de distinguir y pueden separarse fácilmente en dimensiones pequeñas, sin embargo, los no lineales son más difíciles de detectar y por lo tanto no son fáciles de separar, estos patrones necesitan más manipulación que los lineales para poder separarlos de manera más sencilla. La principal función de la SVM es la construcción de un hiperplano óptimo que pueda separar y clasificar patrones lineales y no lineales (Pradhan, 2012). Un hiperplano óptimo es escogido de un conjunto de hiperplanos, maximizando el margen de clasificación de patrones. Para problemas no lineales, se utilizan funciones kernel para mapear los

datos a un espacio de más alta dimensión para convertirlo en un problema lineal (Evgeniou, T., Pontil, M., 2001).

K-vecinos más cercanos (KNN)

K vecinos más cercanos (KNN) es uno de los diversos algoritmos populares para la clasificación supervisada de datos mediante el aprendizaje automático computacional. La fase de entrenamiento consiste en la recopilación de datos de los pares patrón-entrenamiento; asimismo, la fase de clasificación se dedica a definir la similitud de la distancia con respecto a todos los patrones anteriores. Después, se toman los K número de valores anteriores con los que se obtuvieron menor distancia, pero mayor semejanza, es así como se adjuntan a sus clases indicadas. A continuación, se contabiliza el número de veces en el que cada clase aparece en estos patrones vecinos. Por último, el patrón de entrada pertenece a aquella clase que haya obtenido el menor valor de distancia, es decir, el mayor valor de similitud

Árboles de Decisión (DT)

Los árboles de decisión son clasificadores expresados como una partición recursiva del espacio de instancias. Los árboles de decisión consisten de nodos los cuales forman un árbol; un nodo con conexiones salientes se les conoce como nodo interno, mientras que los nodos finales se les conoce como nodos hojas (Rokach, 2005). Cada nodo interno se divide en dos o más de acuerdo a una función específica en base a los datos de entrada. Cada uno de los nodos hojas es clasificado de acuerdo a la clase más apropiada de acuerdo a sus atributos. Los árboles de decisión crean una ramificación de decisiones en base al conjunto de datos de entrenamiento, y el principal objetivo es encontrar la ramificación óptima por medio de la minimización del error de generalización. Esta optimización se realiza por medio de funciones las cuales ayudan a definir un criterio de decisión la cual mide la cantidad de información en cada una de las ramificaciones.

Base de datos

En este estudio se utilizaron las versiones DR16 (Data Release 16) y DR17 (Data Release 17) del Estudio Digital del Cielo Sloan (SDSS, por sus siglas en inglés), las cuales están compuestas por 100,000 observaciones del espacio, cada una. Cada observación está descrita por 17 características y una etiqueta que la identifica como estrella, galaxia o quásar. Las características se pueden observar en la Tabla 1

Tabla 1. Características de las observaciones del SDSS DR16 y DR17

Característica	Descripción
obj_ID	Identificador del objeto, valor único que lo identifica en el catálogo
alpha	Ángulo de ascensión recto
delta	Ángulo de declinación
u	Filtro ultravioleta en el sistema de fotometría
g	Filtro verde en el sistema de fotometría
r	Filtro rojo en el sistema de fotometría
i	Filtro de infrarrojo cercano en el sistema de fotometría
z	Filtro infrarrojo en el sistema de fotometría
run_ID	Número de ejecución utilizado para identificar el escaneo específico
rerun_ID	Número que especifica cómo fue procesada la imagen
cam_col	Número que identifica la línea de exploración

field_ID	Número de campo
spec_obj	Clave única utilizada para objetos espectroscópicos ópticos
class	Clase (galaxia, estrella o quásar)
redshift	Corrimiento al rojo
plate	Número de plato
MJD	Fecha Juliana modificada, utilizada para indicar cuando una pieza del SDSS fue tomada
fiber_ID	Identificación de la fibra que apuntó la luz al plano focal en cada observación

SDSS-DR16 (Data Release 16)

Este conjunto de datos es el cuarto lanzamiento de la cuarta fase del Estudio Digital del Cielo Sloan. El DR16 contiene observaciones del cielo en el mes de agosto de 2018. El número de observaciones es 100,000, de las cuales, 38,096 son estrellas, 51,323 son galaxias y 10,581 son quásares. En la Figura 2a se presenta esta distribución.

SDSS-DR17 (Data Release 17)

Este conjunto de datos es el lanzamiento final de la cuarta fase del Estudio Digital del Cielo Sloan. El DR17 contiene observaciones del cielo en el mes de enero de 2021. El número de observaciones es 100,000, de las cuales, 21,594 son estrellas, 59,445 son galaxias y 18,961 son quásares. En la Figura 2b se presenta esta distribución.

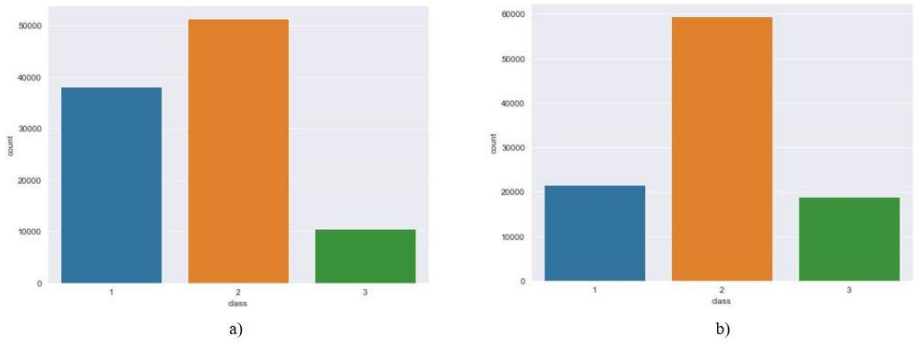

Figura 2. Distribución de frecuencia de observaciones del a) SDSS-DR16 y b) SDSS-DR17. La clase 1 corresponde a estrella, la clase 2 a galaxia y la clase 3 a quásar.

Resultados Experimentales

En esta sección se presentan los resultados obtenidos con cada uno de los algoritmos utilizados en este estudio. El conjunto de datos se dividió en dos subconjuntos, el de entrenamiento (70%) y el de validación (30%).

Regresión Logística

Debido a que el número de muestras de las diferentes clases no es igual, se utilizó como hiper parámetro la ponderación de clases, la cual consiste en ajustar pesos de manera inversa a la frecuencia de las clases:

$$w_i = \frac{N}{NC \times NCi}$$ **Ecuación (1)**

Donde

w_i: Ponderación de la clase i
N: Número total de muestras
NC: Número de clases
NCi: Número de muestras que corresponden a la clase i

El tiempo de entrenamiento de la regresión logística fue de 2.01 segundos utilizando 70,000 datos de entrenamiento y 30,000 de prueba, tanto para el DR16 como para el DR17. Los resultados de la evaluación se presentan en la Tabla 2. Además, en la Figura 4 se presenta la matriz de confusión del resultado de la clasificación.

Tabla 2. Evaluación de la regresión logística utilizando DR16 y DR17

Clase	Precisión	Recall	F1-score
SDSS-DR16			
Estrella	0.97	1.00	0.99
Galaxia	0.99	0.97	0.98
Quásar	0.96	0.96	0.96
Promedio	0.97	0.98	0.98
Exactitud	0.98		
SDSS-DR17			
Estrella	0.95	1	0.97
Galaxia	0.96	0.97	0.96
Quásar	0.95	0.88	0.91
Promedio	0.95	0.95	0.95
Exactitud	0.96		

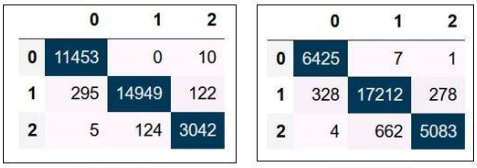

Figura 4. Matrices de confusión de la clasificación realizada por el algoritmo de regresión logística. La clase 0, 1 y 2 son los índices de las clases estrella, galaxia y quásar, respectivamente.

En una matriz de confusión, las columnas son las respuestas correctas mientras que cada renglón son las predicciones de la red. En la Figura 4 se puede observar en las matrices de confusión del SDSS-DR16 y del SDSS-DR17, que la diagonal presenta una mayor densidad, lo que quiere decir que el algoritmo realiza muy bien la clasificación. Sin embargo, en la matriz de confusión del SDSS-DR16, en la columna 0, renglón 1, hay 295 muestras que en realidad son estrellas (columna, clase 0) pero el algoritmo predijo que eran galaxias (renglón, clase 1).

Máquina de Soporte Vectorial (SVM)

Para el entrenamiento se utilizó un valor de regularización de $c = 1$ y se hicieron experimentos con diferentes tipos de kernel, como, por ejemplo, lineal, polinomial, función de base radial y sigmoide, siendo el kernel lineal el que mejor obtuvo resultados. En la Tabla 3 se presentan los

resultados promedio de la evaluación de la SVM y en la Figura 5 sus respectivas matrices de confusión.

Tabla 3. Evaluación de SVM con diferentes tipos de kernels utilizando DR16 y DR17

Kernel	Precisión Promedio	Recall Promedio	F1-score Promedio	Exactitud Promedio
SDSS-DR16				
Lineal	**0.98**	**0.98**	**0.98**	**0.99**
Polinomial	0.94	0.89	0.91	0.90
Función de base radial	**0.98**	0.97	**0.98**	0.98
Sigmoide	0.73	0.72	0.72	0.79
SDSS-DR17				
Lineal	**0.96**	**0.95**	**0.96**	**0.96**
Polinomial	0.53	0.34	0.25	0.60
Función de base radial	0.94	0.92	0.93	0.93
Sigmoide	0.91	0.86	0.88	0.89

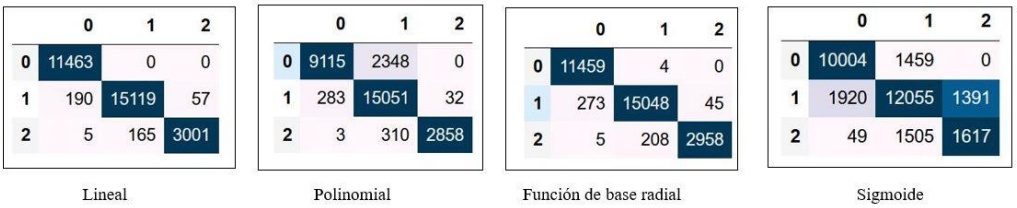

Figura 5. Matrices de confusión de la SVM con sus respectivos kernels utilizando el SDSS-DR16.

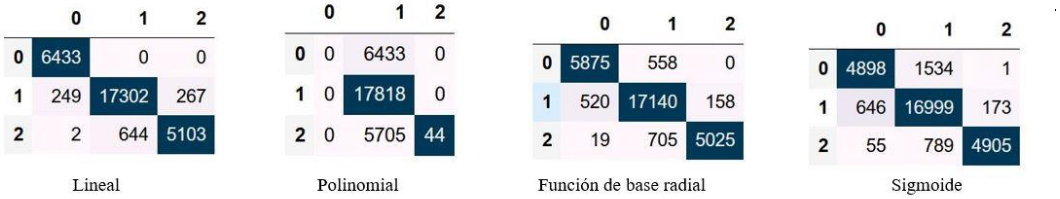

Figura 6. Matrices de confusión de la SVM con sus respectivos kernels utilizando el SDSS-DR17.

K Vecinos más Cercanos (KNN)

El algoritmo KNN se implementó utilizando un valor de $k = 5$ y utilizando la distancia Minkowski con $p = 2$ (la cual también se le conoce como distancia euclidiana) debido a que fueron los hiper parámetros que obtuvieron mejores resultados. La distancia Minkowski está definida como:

$$D(X,Y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$ **Ecuación (2)**

Donde
X : Es un vector de n dimensiones
Y: Es un vector de n dimensiones

Para la distancia euclidiana, hacemos $p = 2$:

$$D(X,Y) = (\sum_{i=1}^{n}|x_i - y_i|^2)^{\frac{1}{2}}$$ **Ecuación (3)**

Los resultados obtenidos por KNN se muestra en la Tabla 4 y en la Figura 7.

Tabla 4. Evaluación de KNN utilizando el DR16 y DR17

Clase	Precisión	Recall	F1-score
	SDSS-DR16		
Estrella	0.97	0.99	0.98
Galaxia	0.98	0.97	0.98
Quásar	0.98	0.93	0.96
Promedio	0.97	0.97	0.97
Exactitud	0.97		
	SDSS-DR17		
Estrella	0.93	0.94	0.93
Galaxia	0.95	0.96	0.96
Quásar	0.97	0.92	0.94
Promedio	0.95	0.95	0.95
Exactitud	0.95		

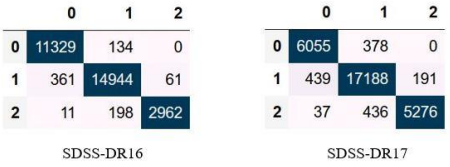

Figura 7. Matrices de confusión de KNN utilizando el SDSS-DR17.

Árboles de Decisión (DT)

Para el entrenamiento de los árboles de decisión se hicieron varios experimentos en donde se concluyó que la máxima profundidad del árbol debería de ser de 5 para obtener los mejores resultados. La función para medir la cantidad de información en cada una de las decisiones de cada nodo fue la entropía. En la Tabla 5 y en la Figura 8 se muestran los resultados de este algoritmo.

Tabla 5. Evaluación del árbol de decisión utilizando el DR16 y DR17

Clase	Precisión	Recall	F1-score
	SDSS-DR16		
Estrella	1.00	1.00	1.00
Galaxia	0.99	0.99	0.99
Quásar	0.98	0.94	0.96
Promedio	0.99	0.99	0.98
Exactitud	0.99		

SDSS-DR17			
Estrella	1.00	1.00	1.00
Galaxia	0.97	0.98	0.98
Quásar	0.95	0.93	0.94
Promedio	0.97	0.98	0.97
Exactitud	0.98		

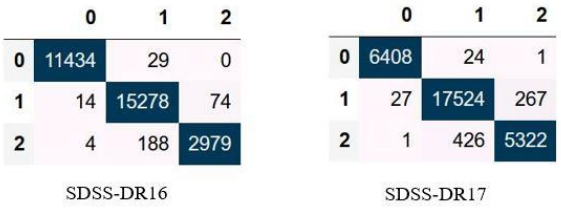

Figura 8. Matrices de confusión de los árboles de decisión utilizando el SDSS-DR16 y SDSS-DR17.

Algo a destacar de los resultados obtenidos por el árbol de decisión es que se truncó su crecimiento con una profundidad máxima de 10 para reducir el sobre ajuste y aún así los resultados son muy buenos.

Tabla 6. Comparación de algoritmos

Clase	Precisión	Recall	F1-score	Exactitud
SDSS-DR16				
Regresión Logística	0.97	**0.98**	**0.98**	0.98
SVM	0.98	**0.98**	**0.98**	**0.99**
KNN	0.98	0.96	0.97	0.97
Árboles de decisión	**0.99**	**0.98**	**0.98**	**0.99**
SDSS-DR17				
Regresión Logística	0.96	0.95	0.95	0.96
SVM	0.96	0.95	0.96	0.96
KNN	0.96	0.94	0.94	0.95
Árboles de decisión	**0.97**	**0.97**	**0.97**	**0.98**

Conclusiones

La clasificación de objetos celestes como las estrellas, galaxias o y quásares permite a los astrónomos poder entender los orígenes del universo; tradicionalmente se utilizaba la información morfológica, sin embargo, para objetos muy lejanos se convierte en un problema, ya que las imágenes obtenidas son de muy baja resolución. Actualmente se utiliza la información espectral para la caracterización de estos cuerpos. Debido a que la cantidad de observaciones realizadas por telescopios es enorme, clasificar estos cuerpos estelares se convierte en una tarea muy difícil y muchas veces imposible. En este artículo analizamos cuatro tipos de algoritmos de aprendizaje automático como por ejemplo la regresión logística, las máquinas de soporte vectorial, las K-vecinos más cercanos y los árboles de decisión. Se utilizaron dos catálogos, el SDSS-DR16 y el SDSS-DR17 y para la evaluación se utilizaron métricas como por ejemplo la precisión, recall, métrica F1 y la exactitud. Los resultados muestran que los algoritmos de machine learning

encuentran patrones dentro de los datos en base a la correlación de variables, en donde el mejor algoritmo analizado fue el árbol de decisión. Algo a destacar de los experimentos, es que el árbol de decisión utilizado en este trabajo fue truncado, esto para prevenir el sobre ajuste.

Agradecimientos

Los autores agradecen al Tecnológico de Monterrey campus Cuernavaca por el apoyo brindado durante la investigación y escritura de este artículo.

Referencias bibliográficas

Bai, Y., Liu, J., Wang, S., Yang, F. (2018) Machine learning applied to star-galaxy-QSO classification and stellar effective temperature regression. *The Astronomical Journal*, **157**(1), 1-9.

Breiman, L. (2001) Random forest. *Machine Learning*, **45**, 5-32.

Clarke, O., Scaife, A., Greenhalgh, R., Griguta, V. (2020) Identifying galaxies, quasars, and stars with machine learning. A new catalogue of classifications for 111 million SDSS sources without spectra. *Astronomy & Astrophysics*, **639**, 1-29.

Costa, M., Sampedro, L., Molino, A., Xavier, H., et al. (2018) The S-PLUS: a star/galaxy classification based on a machine learning approach. arXiv, 2019, desde: https://arxiv.org/abs/1909.08626.

Evgeniou T., Pontil, M. (2001) Support vector machines: theory and applications. Machine Learning and Its Applications. En Paliouras, G., Karkaletsis, V., Spyropoulos, C.D. (eds), *Lecture Notes in Computer Sciences*, Springer, Berlín, Heidelberg, 249-257.

Fadely, R., Hogg, D., Willman, B. (2012) Star-galaxy classification in multi-band optical imaging. *The Astrophysical Journal*, **760**(15), 1-10.

Gaia Collaboration et al. (2016) The Gaia mission. *Astronomy & Astrophysics*, **595**, 1-36.

Gal, R., Carvalho, R, Odewahn, S., Djorgovski, S., Mahabal, A, Brunner, R, Lopes, P. (2004) The digitized second palomar observatory sky survey (DPOSS). II. Photometric calibration. *The Astronomical Journal*, **128**, 3082-3091

Kim, E., Brunner, R. (2016) Star galaxy classification using deep convolutional neural networks. *Monthly Notices of the Royal Astronomy Society*. **464**, 4463-4475.

Krakowski, T., Malek, K., Bilicki, M., Pollo, A., Kurcz, A., Krupa, M. (2016) Machine-learning identification of galaxies in the WISE x SuperCOSMOS all-sky catalogue. *Astronomy & Astrophysics*, **596**(39), 1-11.

Pradhan, A. (2012) Support vector machine – a survey. *International Journal of Emerging Technology and Advanced Engineering*, **2**(8), 82-85.

Robertson, B., Ellis, R., Furlanetto, S., Dunlop, J. (2015) Cosmic reionization and early star-forming galaxies: a joint analysis of new constraints from Planck and the Hubble Space Telescope. *The Astrophysical Journal Letters*, **802**, 1-5.

Rokach, L., Maimon, O. (2005) Decision trees. En Maimon, O., Rokach, L. (eds). *Data Mining and Knowledge Discovery Handbook*, Springer, Boston, MA, 165-192.

Solarz, A., Pollo, A., Takeuchi, T., Pepiak, A., Matsuhara, H., Wada, T., Oyabu, S., Takagi, T, Goto, T., Ohyama, Y., Pearson, C, Hanami, H, Ishikagi, T. (2012) Star-galaxy separation in the AKARI NEP deep field. *Astronomy & Astrophysics*, **541**, 1-8.

Vasconcellos, E., Carvalho, R., Gal, R., Labarbera, F. (2011) Decision tree classifiers for star/galaxy separation. *The Astronomical Journal*, **141**(189), 1-12.

Xiao, W., Jin, Y. (2021) Classification of star/galaxy/QSO and star spectral types from LAMOST data release 5 with machine learning approaches, *Chinese Journal of Physics*, **69**, 303-311.

Zou, X., Hu, Y., Tian, Z., Shen, K. (2019) Logistic regression model optimization and case analysis. *IEEE International Conference on Computer Science and Network Technology*, Dalian, China, Oct 19, 2019.