Clasificación de objetos astronómicos con redes neuronales

Brayhan Pérez Rendon*
Universidad de Antioquia
(FACOM)
(Dated: June 17, 2025)

I. INTRODUCCIÓN

En este documento se propone una posible solución a un desafío que ha acompañado tanto a la astronomía moderna como a sus raíces históricas. El proyecto se fundamenta en la competencia de Kaggle titulada "The Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC)". El Legacy Survey of Space and Time (LSST), que se llevará a cabo desde el Observatorio Vera C. Rubin en Cerro Pachón, Chile, constituye una misión astronómica ambiciosa de diez años. Emplea un telescopio Simonyi de 8,4m con diseño de tres espejos y cuenta con la cámara digital más grande jamás construida (3200 megapíxeles). Su propósito es capturar imágenes de toda la parte visible del cielo del hemisferio sur cada pocos días, generando una "película del universo" en movimiento. Entre sus objetivos principales se encuentran investigar la materia y energía oscuras, catalogar objetos del Sistema Solar (incluyendo asteroides potencialmente peligrosos), monitorear fenómenos transitorios como supernovas y estallidos de rayos gamma, y mapear la estructura de la Vía Láctea .

Esta misión ha generado un conjunto de datos de aproximadamente 40GB con objetos astronómicos relevantes, que incluye tanto datos etiquetados para el entrenamiento del modelo como datos sin clasificar para evaluar su desempeño en escenarios de gran volumen de información. El objetivo es que la comunidad de Kaggle, utilizando la métrica Log-loss, desarrolle un modelo de Machine Learning capaz de clasificar dichos objetos [1]. Entre sus características se incluyen series temporales de flujo de fotones captados por el telescopio en la banda visible, observaciones en seis filtros distintos dentro de esa región del espectro, correcciones por extinción causada por polvo intergaláctico, así como información de distancia y redshift para objetos extragalácticos. Esta clasificación abarca una amplia variedad de fuentes astronómicas, muchas de las cuales presentan variabilidad temporal, como pulsares, cuásares, sistemas binarios o supernovas, cuya identificación específica no siempre se detalla o explica en el documento principal. Las propiedades de estos objetos se analizarán y se interpretarán en los resultados del modelo propuesto.

El problema a resolver radica en la clasificación automática y efectiva de fuentes astronómicas, una tarea que, de hacerse manualmente, podría tomar meses para un conjunto reducido de objetos. Aquí el aprendizaje automático destaca por su rapidez y precisión al procesar enormes volúmenes de eventos. Esto resulta crucial, pues la capacidad de manejar un caudal de datos que de otro modo sería inviable permite realizar un análisis profundo del comportamiento, la dinámica, las mediciones y la distribución del universo. A su vez, esta posibilidad allana el camino hacia la construcción de una física unificadora capaz de explicar diversos fenómenos astronómicos, tal como indican los objetivos del LSST.

II. MÉTODOS

En esta sección se describen el tratamiento de los datos y el modelo empleado para abordar el problema. La discusión de los resultados y la configuración de los objetos astronómicos se presentará en la sección siguiente; aquí nos enfocamos únicamente en exponer cómo se procesan los datos y la lógica subyacente al modelo de aprendizaje automático utilizado.

A. Modelo

Tras analizar los resultados del concurso iniciada en 2018 y los enfoques de los equipos ganadores, se observa que la mayoría optó por la sencillez y la flexibilidad de modelos basados en árboles de decisión, tanto simples como ensamblados. Por ello, decidí explorar el comportamiento del problema usando redes neuronales.

Las redes neuronales, si bien comparten con los árboles de decisión la capacidad de segmentar el espacio de características, se caracterizan por identificar patrones mediante operaciones lineales y no lineales encadenadas a lo largo de múltiples capas. Cada ruta entre capas busca transformar las entradas en representaciones que faciliten la clasificación en una de las clases objetivo, incluso cuando no se dispone de información previa explícita sobre la estructura de dichos patrones.

En nuestro caso, se trata de un problema de clasificación supervisada con 14 clases distintas, cada una correspondiente a un tipo de evento u objeto astronómico de interés (por ejemplo, cuásares, estrellas variables, pulsares, supernovas y otros fenómenos con variabilidad temporal). Contamos con dos conjuntos de datos principales:

1. **Metadata**: contiene información estática de cada objeto, asignado con un identificador interno de la

* Also at Instituto de Física, Maestría en Física

misión. Incluye coordenadas eclípticas y galácticas, redshift espectral y fotométrico, distancia estimada, extinción debida a absorción de la luz por polvo y gas de la Vía Láctea, y el campo "target" que indica el tipo de objeto mediante un número entero.

2. **Series de flujo temporal**: registra el flujo de fotones captado por el Simonyi Survey Telescope a lo largo del tiempo. Las tomas pueden ser diarias o semanales hasta que el objeto deja de ser visible; en muchos casos se efectúan segundas o terceras rondas de recolección, lo que genera intervalos sin mediciones conocidas, aunque en general la evolución del comportamiento queda bien documentada.

Para abordar el problema, se emplea MLPClassifier de la librería scikit-learn: un clasificador basado en un perceptrón multicapa que aprende relaciones no lineales entre características y etiquetas, entrenándose mediante retropropagación y optimizando la función de pérdida de entropía cruzada (log-loss). Como configuración inicial, se propone una arquitectura de capas ocultas con 128, 64 y 32 neuronas, respectivamente. Tras entrenar con los datos de entrenamiento y evaluar sobre el conjunto de validación (o test), se examinará la existencia de underfitting (precisión baja tanto en entrenamiento como en validación) o sobreajuste (precisión alta en entrenamiento pero deficiente en validación). Según estos resultados, se considerará profundizar (más capas o neuronas) en caso de underfitting o aplicar regularización y/o reducir la complejidad si se detecta sobreajuste. El objetivo fundacional es que el modelo ofrezca resultados estables y consistentes, aunque no alcance una exactitud perfecta.

Finalmente, y siguiendo el formato requerido por el reto de Kaggle, la salida consistirá en una matriz donde cada fila contiene el ID del objeto y las probabilidades estimadas de pertenencia a cada una de las 14 clases disponibles.

B. Tratamiento de datos

El buen uso de los datos disponibles facilita los resultados de cualquier modelo de ML, aquí se va a abordar el tratamiento de los datos preliminar. Los datos como la redshift espectral del objeto, redshift fotométrico, distancia y Extensión son importantes características puntuales, que en algunos casos es muy difícil de obtener. El redshift espectral y fotométrico es información altamente valiosa porque da características del objeto, si este está fuera de la galaxia, si este está muy lejos (información vital para descartar eventos poco energéticos) y además si este tiene rotación o está en movimiento relativo con nosotros y cual es su sentido; la segunda propiedad importantes es la distancia directa, descartando objetos según su cercanía, y la extensión, hace que se pueda predecir la energía real del evento u objeto, ya que el espacio

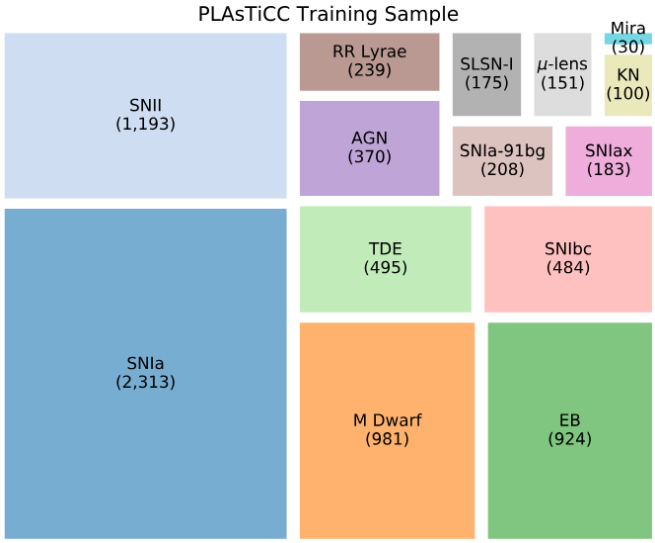

FIG. 1: Los 14 tipos de objetos astronómicos clasificado, imagen tomada de [1]

intergaláctico podría absorber parte de la energía de la luz emitida, confundiendo a la red; aunque no se tiene información directa en los targets de que tipo de cuerpos astronómicos, ya que no se proporciona, se puede deducir la importancia de cada parámetros debido a la evidencia y previo estudio en fenómenos relacionados con estas propiedades.

Las propiedades mas importantes que se pueden obtener del objeto, es como se distribuye el flujo de luz a lo largo del tiempo, esto es imprescindible para facilitar el reconocimiento de diversos objetos, Pero este tipo de datos están repartidos en una serie de tiempo irregular, con datos de diferentes tamaños y tiempos, esto hace difícil dar directamente los datos a la red neuronal y es por esto por lo que se usaran propiedades estadísticas de la distribución del flujo a lo largo del tiempo. Se usarán las siguientes propiedades:

- **El promedio**: Útil para ver le equilibrio estadístico de la distribución.
- **La desviación estándar**: para saber el ancho de este y descartar datos por fuera de lo común.
- **Skewness**: Permite medir la simetría de la distribución
- **Kurtosis**: Permite medir que tan gruesas con las colas de la distribución

Esto se hace en cada filtro y en la combinación de todas estas, más los puntos mínimos y máximos en cada una de estas bandas, estas bandas son lo más relevante y lleno de información de tiene la distribución, ya que la comparación entre cada filtro da información espectral del objeto, esta característica pueda dar una muy buena idea que tipo de objeto es y la energía del evento en cuestión, esta relación se llama Color en la astronomía

y se hace con la resta de picos en dos filtros, el resultado de esto da información sobre la banda dominante del objeto. Como los datos que se tomaron está en el visible, cuanto más dominen los flujos en filtros azulados, más energético será el evento, o bien puede dar información sobre temperatura de radiación, y demás características que no se hace de manera previa pero la red se encargara de encontrar estas relaciones.

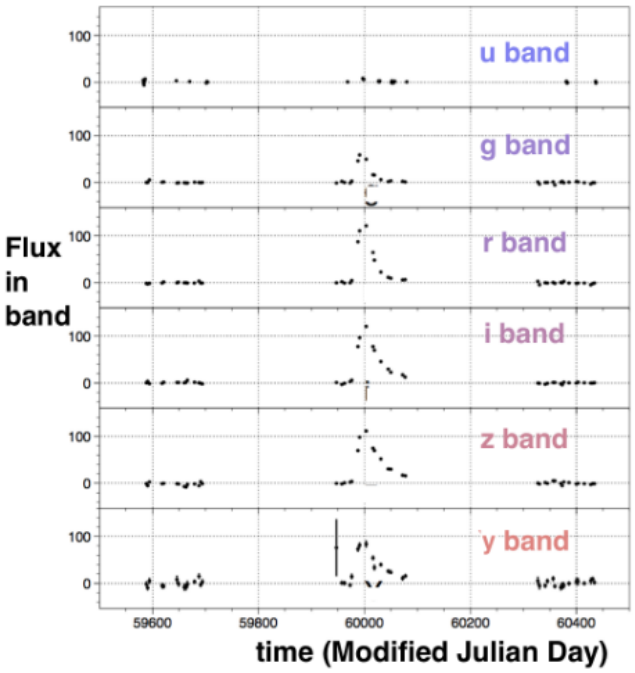

FIG. 2: Ejemplo de uno de los objetos del set de entrenamiento, y como cambia según el el filtro, evidenciando la importancia de trata el filtro parte por parte

Y por último el número de detecciones en cada banda, que indica la detección del objeto dando pistas de la función espectral del objeto. Como medio comparativo y experimental, como primera parte se hace uso únicamente de los datos extraídos por la curva de luz en el tiempo y luego se añade las características adicionales, No se hace uso de las coordenadas, no son muy relevantes teniendo en algunos casos la distancia.

III. RESULTADO Y DISCUSIÓN

El primer entrenamiento, empleando únicamente las propiedades derivadas de la curva de luz, arroja resultados interesantes: a pesar de la simplicidad de la arquitectura, se obtiene un score general de 0.765, lo que demuestra la viabilidad del modelo como clasificador. Al añadir características extra en un segundo experimento, el score general apenas varía (0.764), pero al desglosar el desempeño por clase se observa que la detección de ciertos tipos de objetos mejora notablemente, en otros casos se mantiene igual y en algunos empeora. Estos hallazgos proporcionan pistas sobre qué categorías prioriza el modelo y cuáles pueden requerir ajustes específicos. En conjunto, aunque la precisión global haya disminuido ligeramente, esta segunda aproximación aporta información valiosa y supone una mejora en la comprensión del comportamiento del clasificador.

Clase	Precisión	Recall
6	0.79	0.77
15	0.66	0.58
16	0.94	0.95
42	0.66	0.56
52	0.29	0.11
53	1.00	1.00
62	0.44	0.52
64	0.61	0.70
65	0.96	0.92
67	0.29	0.26
88	0.92	0.96
90	0.77	0.86
92	0.87	1.00
95	0.91	0.83

TABLE I: Precisión y Recall por clase, para el entrenamiento del modelo sin datos extra, es decir, solo se entrena con propiedades de la curva de luz

Clase	Precisión	Recall
6	0.93	0.93
15	0.72	0.59
16	0.96	0.97
42	0.58	0.45
52	0.00	0.00
53	1.00	1.00
62	0.33	0.51
64	0.68	0.75
65	0.97	0.97
67	0.47	0.36
88	0.99	0.99
90	0.77	0.88
92	1.00	0.92
95	0.88	0.83

TABLE II: Precisión y Recall por clase, con el entrenamiento del modelo con datos extra, como la extensión, redshift, etc

Al centrarnos en el target clasificado como 52, observamos que se trata de un evento astronómico de alta energía que aparece en tres detecciones separadas por aproximadamente 230 días. Para abordar este caso, se planteó realizar un análisis por detección: es decir, dividir la serie de tiempo en tres secciones correspondientes a cada aparición y aplicar en cada segmento las operaciones estadísticas previamente definidas. De este modo, se aprovecha el margen temporal acotado de cada suceso, facilitando que la red neuronal aprenda patrones específicos de cada detección y compararlos entre sí.

Con una puntuación global de 0,53 y sin mejoras en nuestro objetivo para el target 52 (precisión 0,07 y re-

Clase	Precisión	Recall	Cambio
6	0.79 → 0.93	0.77 → 0.93	Mejora
15	0.66 → 0.72	0.58 → 0.59	Mejora
16	0.94 → 0.96	0.95 → 0.97	Mejora
42	0.66 → 0.58	0.56 → 0.45	Empeora
52	0.29 → 0.00	0.11 → 0.00	Empeora
53	1.00 → 1.00	1.00 → 1.00	Igual
62	0.44 → 0.33	0.52 → 0.51	Empeora
64	0.61 → 0.68	0.70 → 0.75	Mejora
65	0.96 → 0.97	0.92 → 0.97	Mejora
67	0.29 → 0.47	0.26 → 0.36	Mejora
88	0.92 → 0.99	0.96 → 0.99	Mejora
90	0.77 → 0.77	0.86 → 0.88	Mejora parcial
92	0.87 → 1.00	1.00 → 0.92	Cambio mixto
95	0.91 → 0.88	0.83 → 0.83	Empeora parcial

TABLE III: Comparación de precisión y recall entre la Tabla 2 y la Tabla 1

call 0,03), queda claro que este tratamiento de datos no aporta valor al problema. Presenta varias desventajas: en primer lugar, reduce el tamaño del conjunto de entrenamiento; en segundo lugar, compromete la relevancia de las características seleccionadas; y, además, implica una implementación más costosa tanto en desarrollo como en recursos computacionales.

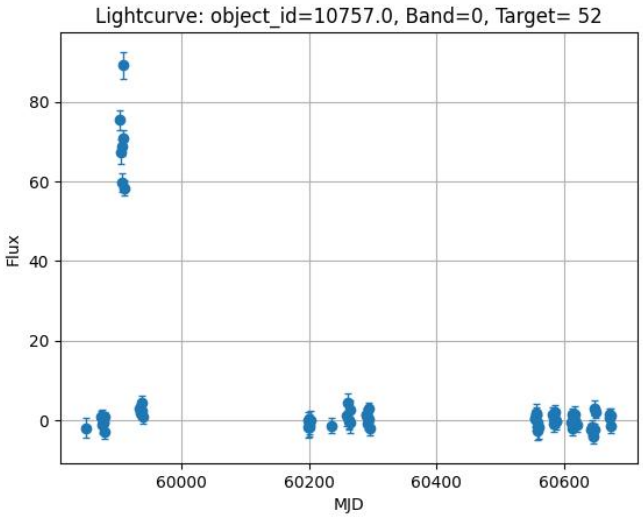

FIG. 3: Objeto del tipo 52, por su comportamiento deduzco que pueda ser alguno de los tipos de supernovas, por su muy tenue duración

Se planteó generar un texto (text) con los datos proporcionados por Kaggle, pero surgieron limitaciones de memoria RAM y tiempos de procesamiento al manejar los 20GB necesarios para que la plataforma pueda evaluar y validar nuestro modelo.

IV. CONCLUSIÓN

Aunque únicamente con el entrenamiento se pudieron observar ciertos comportamientos y extraer utilidades específicas —por ejemplo, un mejor entendimiento de la naturaleza de los datos—, en combinación con un preprocesado adecuado pueden obtenerse resultados notables.

Es evidente que este tipo de modelo resulta difícilmente generalizable a todos los eventos astronómicos presentes en el conjunto de datos; en cambio, se maneja de forma más eficiente si se aborda por categorías de eventos y se clasifica según el comportamiento particular de cada tipo.

También es prioritario verificar el desempeño del modelo con datos de prueba reales, pues ello permitirá detectar sobreajuste (overfitting) o subajuste (underfitting) y ajustar la implementación en consecuencia. Por otro lado, sería deseable entrenar un modelo de mayor complejidad, pero los recursos disponibles en este proyecto no lo permiten. Al probar una cuarta capa con 32 neuronas, el rendimiento general mejoró apenas un 1 %, lo cual sugiere que se requiere un estudio más profundo de la arquitectura.

Cuando se incorporan datos adicionales, el modelo demuestra alta eficacia para cierto tipo de casos, tal como sugiere [2], lo que indica que podría ser viable para clasificaciones astronómicas específicas.

[1] R. Hložek, K. A. Ponder, A. I. Malz, M. Dai, G. Narayan, E. E. O. Ishida, T. A. Jr, A. Bahmanyar, R. Biswas, L. Galbany, S. W. Jha, D. O. Jones, R. Kessler, M. Lochner, A. A. Mahabal, K. S. Mandel, J. R. Martínez-Galarza, J. D. McEwen, D. Muthukrishna, H. V. Peiris, C. M. Peters, and C. N. Setzer, Results of the photometric lsst astronomical time-series classification challenge (plasticc) (2020), arXiv:2012.12392 [astro-ph.IM].

[2] A. D'Isanto, S. Cavuoti, M. Brescia, C. Donalek, G. Longo, G. Riccio, and S. G. Djorgovski, An analysis of feature relevance in the classification of astronomical transients with machine learning methods, Monthly Notices of the Royal Astronomical Society **457**, 3119 (2016), https://academic.oup.com/mnras/article-

pdf/457/3/3119/8001547/stw157.pdf.
[3] K. Boone, Avocado, `https://github.com/kboone/avocado` (2021), accessed: 2025-06-17.
[4] H. Qu, M. Sako, A. Möller, and C. Doux, Scone: Supernova classification with a convolutional neural network, The Astronomical Journal **162**, 67 (2021).