Aplicación de métodos supervisados en la clasificación de objetos celestes

Luis David Huante García
Tecnologías para la Información en Ciencias
UNAM ENES Morelia
luisdhuante@gmail.com

Figure 1: Un hoyo negro interrumpe la trayectoria de una estrella. NASA, JPL-Caltech

ABSTRACT

Este proyecto pone en prática diversos modelos de aprendizaje automático, incluyendo Gaussian Naive Bayes, K-Nearest Neighbors, Decision Tree, Logistic Regression, Support Vector Machine y un ensamble basado en Voting Classifier, para clasificar objetos celestes como galaxias, estrellas y quásares. Utilizando un conjunto de datos que incorpora características fotométricas y espectrales, se llevó a cabo un análisis exploratorio, seguido de preprocesamiento y una optimización de hiperparámetros vía GridSearchCV. Los resultados muestran que el modelo de ensamble superó a los individuales en precisión y F1-Score, demostrando la efectividad de las técnicas de ensamble para la mejora en las clasificaciones.

KEYWORDS

clasificación, astronomía, estrellas, galaxias, cuásares, aprendizaje, ensamble, algoritmos, datos, predicción.

ACM Reference Format:
Luis David Huante García. 2024. Aplicación de métodos supervisados en la clasificación de objetos celestes. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCCIÓN

La clasificación de objetos celestes es una herramienta muy importante en el campo de la astronomía, permitiendo a los científicos comprender mejor la composición y la estructura del universo. Este proyecto se centra en la aplicación de técnicas de aprendizaje automático supervisado para la clasificación de estrellas, galaxias y cuásares utilizando datos obtenidos del Sloan Digital Sky Survey (SDSS). Con este proyecto, se busca explorar la capacidad de

diferentes algoritmos al categorizar estos objetos astronómicos basándonos en sus características espectrales.

En el contexto de la astronomía, el aprendizaje automático puede ser particularmente valioso, permitiendo análisis complejos de grandes conjuntos de datos. La SDSS proporciona un recurso extenso con más de 100,000 observaciones, cada una descrita por 17 atributos. [SDS [n. d.]]

En este trabajo, se inmplementó y comparó el rendimiento de varios algoritmos clásicos de aprendizaje automático, incluyendo Naive Bayes, k-Vecinos más cercanos, Árboles de decisión, Regresión logística y Máquinas de soporte vectorial. Además, se desarrolló un método de ensamble que combina las predicciones de estos modelos individuales.

2 QUÁSARES, ESTRELLAS Y GALAXIAS

Antes de adentrarnos en la metodología del proyecto, es importante definir claramente los objetos astronómicos que clasificaremos: estrellas, galaxias y quásares. Cada uno de estos tipos de objetos juega un papel crucial en nuestra comprensión del universo y presenta características únicas que los distinguen:

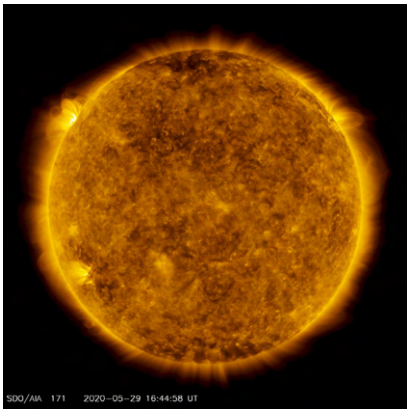

Figure 2: Imagen del Sol captada por el Observatorio de Dinámica Solar de la NASA. NASA/SDO

Estrellas: Las estrellas son enormes esferas de plasma incandescente, mantenidas por la fuerza de la gravedad y alimentadas por reacciones nucleares en sus núcleos. La vida de una estrella comienza en las nebulosas, nubes de gas y polvo, donde las perturbaciones como ondas de choque de supernovas cercanas pueden iniciar el colapso que eventualmente forma nuevas estrellas. A lo largo de su vida, que puede variar desde unos pocos millones hasta más de diez mil millones de años, una estrella pasa por diversas fases, determinadas principalmente por su masa inicial. Estas fases incluyen la secuencia principal, gigante roja, y, eventualmente, dependiendo de su masa, puede terminar como enana blanca, estrella de neutrones o agujero negro. [nas [n. d.]b]

Galaxias: Las galaxias son vastas agrupaciones de estrellas, gas, polvo y materia oscura, unidas por la gravedad. Existen varios tipos de galaxias, incluidas las espirales, elípticas y lenticulares, cada una con características morfológicas distintivas. Las galaxias espirales, como la Vía Láctea, poseen discos rotativos con brazos espirales y un bulbo central; las galaxias elípticas carecen de estructura de

Figure 3: La galaxia de Andrómeda, también conocida como Messier 31, aparece en esta imagen del Wide-field Infrared Survey Explorer (WISE) de la NASA. NASA/JPL-Caltech/UCLA

disco y tienen perfiles de brillo más suavizados; y las lenticulares presentan un disco sin brazos espirales. Las interacciones entre galaxias, que pueden incluir fusiones y tránsitos gravitacionales, juegan un papel importante en su evolución y en la formación de estructuras a gran escala en el universo. [nas [n. d.]a]

Figure 4: Concepto de una galaxia con un cuásar brillante en su centro. NASA, ESA and J. Olmsted (STScI)

Quásares: Los quásares son los núcleos activos de galaxias jóvenes y distantes, alimentados por agujeros negros supermasivos en sus centros. Estos son algunos de los objetos más brillantes del universo, a pesar de su tamaño compacto, comparable al de nuestro sistema solar. Los quásares son particularmente valiosos para los astrónomos debido a su visibilidad a distancias extremas, lo que los convierte en herramientas esenciales para sondear las regiones más remotas del universo y entender la evolución galáctica a lo largo del tiempo cósmico. La intensa actividad en los quásares resulta en una emisión significativa a lo largo de todo el espectro electromagnético, desde ondas de radio hasta rayos X.[nas [n. d.]b]

Ahora se explicarán las variables del dataset del Sloan Digital Sky Survey (SDSS). Estas variables son esenciales para identificar y clasificar objetos astronómicos como los que conforman el dataset. Cada una de ellas aporta información que mejora la precisión del modelo.

3 ¿CUÁL ES CUÁL?

Es crucial entender cómo cada una de las variables contribuye específicamente al proceso de análisis y a cómo cada una distingue a cada cuerpo celeste:

3.0.1 Mediciones en múltiples bandas del espectro (u, g, r, i, z). Estas variables representan las intensidades observadas en cinco bandas espectrales distintas, que van desde el ultravioleta (u) hasta el infrarrojo cercano (z). Cada banda recoge información específica sobre la emisión de luz de los objetos celestes, variando según la temperatura, la composición química, y otras propiedades físicas. Por ejemplo, las estrellas más calientes y jóvenes tienden a emitir más luz en las bandas ultravioletas, mientras que las galaxias con formación estelar activa pueden mostrar un brillo más intenso en las bandas infrarrojas debido al polvo calentado por estrellas jóvenes. Estas diferencias son críticas para clasificar correctamente los objetos celestes, permitiendo distinguir entre estrellas, galaxias y quásares según sus perfiles espectrales.[of Encyclopaedia Britannica 2024]

3.0.2 Redshift (desplazamiento hacia el rojo). El redshift es fundamental para determinar la distancia a la que se encuentra un objeto celeste y, por implicación, su velocidad de alejamiento debido a la expansión del universo. En el contexto de la clasificación, el redshift puede ayudar a diferenciar entre galaxias y quásares, ya que estos últimos generalmente presentan valores de redshift más altos, indicando que están más lejanos y, por lo tanto, pertenecen a las etapas más tempranas del universo. Además, el análisis del redshift permite estudios detallados sobre la distribución de galaxias y la estructura a gran escala del universo. [esa [n. d.]]

3.0.3 Identificadores Obj_ID y Spec_obj_ID. Estos son identificadores únicos para cada observación y para los objetos espectroscópicos, respectivamente. El manejo adecuado de estos identificadores asegura que la información se atribuya al objeto correcto y que se mantenga la integridad del análisis. Por ejemplo, múltiples observaciones del mismo objeto (que pueden ocurrir en diferentes momentos o bajo diferentes condiciones de observación) deben ser correlacionadas correctamente para obtener una comprensión precisa de sus características.

3.0.4 Fiber ID. Esta variable identifica la fibra óptica específica a través de la cual se canalizó la luz del objeto en el telescopio. La precisión en el uso de *Fiber_ID* es vital para la calibración correcta de los datos, ya que variaciones en la posición de la fibra pueden afectar las mediciones.

4 CLASIFICACIÓN

Ahora se detallará el proceso relacionado con los modelos de clasificación empleados en la clasificación de objetos celestes. Esta fase del proyecto abarca desde la selección y optimización de los modelos hasta la evaluación de su desempeño. A través de este segmento, se explicarán las técnicas de análisis exploratorio de datos, limpieza de datos y optimización de hiperparámetros. Además, se discutirán los criterios usados para evaluar cada modelo, incluyendo la elección de métricas específicas como el F1-score y el F2-score, que son esenciales para balancear la precisión y el recall en un contexto donde la precisión de la clasificación es de suma importancia.

4.1 Análisis exploratorio

Se llevó a cabo un análisis exploratorio de datos para comprender la estructura y las características del conjunto de datos. Este análisis inicial incluyó la visualización de distribuciones de características, la evaluación de la correlación entre diferentes variables y la identificación de posibles valores atípicos. Herramientas como histogramas, diagramas de dispersión y mapas de calor fueron utilizadas. Esta etapa fue esencial para asegurar que las características seleccionadas fueran relevantes para los modelos y que cualquier anomalía en los datos no distorsionara los resultados de los modelos. En el análisis del conjunto de datos, utilizamos el método `describe()` de Python para obtener un resumen estadístico que incluye:

- **count**: Número de observaciones no nulas.
- **mean**: Media aritmética de los datos.
- **std**: Desviación estándar, que mide la dispersión.
- **min**: Valor mínimo en cada columna.
- **25%**, **50%** (mediana), y **75%**: Cuartiles que indican la distribución de los datos.
- **max**: Valor máximo en cada columna.

Este método es esencial para comprender la tendencia central, la dispersión y los límites de los rangos de datos, facilitando las decisiones sobre las técnicas de preprocesamiento y análisis posterior. Por ejemplo, en este análisis se observó que el valor mínimo de ciertas columnas era -9999, lo cuál parecía indicar un error en la lectura de las mediciones tomadas o un valor planchado a observaciones erróneas.

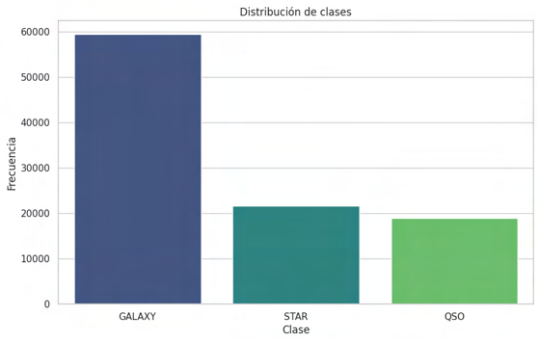

Figure 5: Distribución de clases (Galaxia, Estrella y Quasar en el dataset)

Como parte del análisis exploratorio se creó este gráfico, que ilustra un desbalance significativo en la cantidad de observaciones de cada tipo de objeto astronómico. La predominancia de galaxias sobre estrellas y quásares puede influir en cómo los modelos aprenden a clasificar nuevos datos, posiblemente sesgando los modelos hacia la clase más común. Esto conlleva a la necesidad de considerar técnicas como el ajuste de pesos, sobremuestreo de las clases minoritarias, o submuestreo de la clase mayoritaria para equilibrar el conjunto de datos antes de aplicar algoritmos de clasificación. Además, esta distribución puede tener implicaciones en cómo se interpretan los resultados, y por ello es importante utilizar métricas que puedan dar cuenta del desbalance de clases, como el F1-score

o el F2-score, para una evaluación más justa del rendimiento del modelo.

Figure 6: Matriz de correlación de variables numéricas del dataset

La anterior imagen muestra una matriz de correlación que se generó como parte del análisis exploratorio de los datos. En la matriz, las características 'u' y 'g' muestran una correlación perfecta de 1.00, lo que se espera porque estas magnitudes son medidas fotométricas similares tomadas en bandas espectrales cercanas. Asimismo, 'r' e 'i', y 'i' y 'z' también exhiben correlaciones muy altas. En cambio, características como 'alpha' y 'u' muestran muy poca correlación.

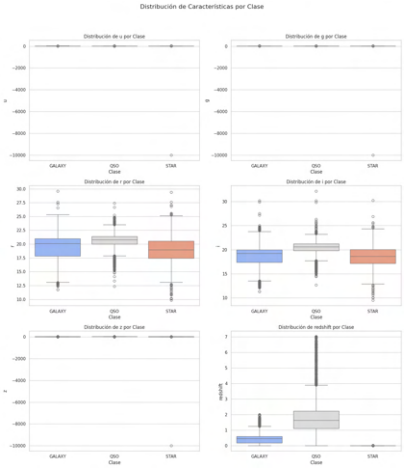

Figure 7: Distribución en gráficos de caja, sesgados por los valores atípicos.

La anterior imagen muestra un conjunto de gráficos de caja (box plots) que representan la distribución de varias características fotométricas (u, g, r, i, z) y el redshift por clase de objeto celeste (Galaxia, QSO, Estrella). Cada gráfico divide las observaciones en las tres categorías y da una representacion visual de la mediana, los cuartiles y los valores atípicos.

Uno de los problemas evidentes en varios de los gráficos es la presencia de valores atípicos extremos, especialmente en las características 'u', 'g' y 'z', donde los valores se extienden hasta cerca de -10000, lo cual es altamente inusual y distorsiona la escala de los gráficos. Esta presencia de valores extremadamente bajos puede ser indicativa de errores de medición o de entrada de datos.

Para solucionar este problema, se realizó un proceso de limpieza de datos que involucró la identificación e imputación de estos valores atípicos. Al remover o ajustar estos valores extremos, los gráficos de caja resultantes muestran una representación más fiel y precisa de la distribución de datos. Este ajuste permite una mejor visualización y comparación entre las clases, además de asegurar que los modelos no sean influenciados negativamente por datos atípicos.

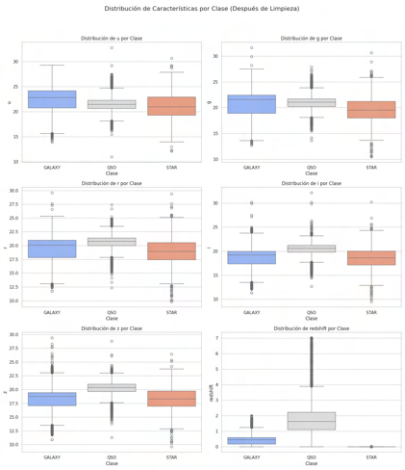

Figure 8: Distribución en gráficos de caja corregida y sin valores atípicos.

Algunas de los insights obtenidos tras el análisis exploratorio fueron los siguientes:

Magnitudes ('u', 'g', 'r', 'i', 'z'):

- **Estrellas**: Tienen generalmente magnitudes más altas (menos negativas, lo que indica mayor brillo) en todas las bandas comparadas con las galaxias y quásares. Esto va acorde con el hecho de que las estrellas son objetos más cercanos y, por lo tanto, aparecen más brillantes en estas bandas.
- **Galaxias**: Presentan una variabilidad considerable en sus magnitudes, pero tienden a ser menos brillantes que las estrellas, lo cual refleja su mayor distancia de la Tierra. La dispersión en las magnitudes también puede indicar una diversidad en el tipo y tamaño de las galaxias observadas.
- **Quásares**: Muestran un rango similar de magnitudes a las galaxias pero con una dispersión ligeramente menor. Los quásares, dado que son núcleos activos, pueden ser extremadamente brillantes.

Redshift:

- **Estrellas**: Los valores de redshift son cercanos a cero, lo que es esperado ya que el redshift mide la velocidad relativa de alejamiento y las estrellas en nuestra galaxia no se alejan a velocidades comparables a las de galaxias o quásares.

- **Galaxias y Quásares**: Ambos muestran valores más altos de redshift, con los quásares típicamente exhibiendo los valores más altos. Esto refleja su gran distancia desde la Tierra y la expansión del universo, que aleja los objetos más distantes a velocidades más altas.

Implicaciones para el modelo:

- **Diferenciación entre clases**: Las estrellas se pueden identificar claramente por su brillo y bajo redshift, mientras que los quásares y galaxias pueden diferenciarse más por sus valores de redshift y variaciones en las magnitudes.
- **Preprocesamiento de características**: Debido a la variabilidad en las escalas de magnitudes y redshift, la normalización o estandarización es importante para asegurar que todas las características contribuyan correctamente al modelo.

4.2 Implementación de modelos

4.2.1 Preprocesamiento. Eliminar o corregir outliers es crucial para el preprocesamiento, ya que mejora la calidad del análisis y asegura que los resultados del modelo sean más robustos y confiables. En este caso específico, los valores atípicos extremos identificados en las variables fueron tratados mediante su reemplazo por la mediana de la respectiva variable. La elección de la mediana como valor de reemplazo se debe a su robustez frente a los outliers, ya que no se ve tan afectada por valores extremos como lo haría la media. En el proceso de análisis del conjunto de datos, una observación importante fue la ausencia de valores nulos.

Los hiperparámetros son una parte fundamental del ajuste de los modelos, y para la selección y optimización de hiperparámetros se utilizó el método GridSearchCV, una herramienta ampliamente utilizada. GridSearchCV realiza una búsqueda exhaustiva sobre un espacio de hiperparámetros especificado, probando todas las combinaciones posibles y evaluando su rendimiento mediante validación cruzada. Este método no solo permite identificar la configuración óptima de hiperparámetros para cada modelo, sino que también ayuda a comparar el rendimiento de los modelos bajo diferentes configuraciones. GridSearchCV facilita significativamente el proceso de ajuste de los modelos, asegurando que se adopten las configuraciones que maximizan la precisión y la eficiencia del modelo final. Esta estrategia es esencial para desarrollar sistemas de clasificación robustos y confiables que puedan operar eficazmente en el complejo y variado entorno de datos astronómicos.

- **Gaussian Naive Bayes:**
  - No requiere ajuste de hiperparámetros, asume independencia entre características.
- **K-Nearest Neighbors (KNN):**
  - **Métrica**: `manhattan` – Utiliza la suma de las diferencias absolutas de las coordenadas.
  - **Número de Vecinos (`n_neighbors`)**: 5 – Número óptimo de vecinos para la clasificación.
  - **Pesos (`weights`)**: `distance` – Los vecinos más cercanos influyen más en el resultado.
- **Decision Tree:**
  - **Profundidad máxima (`max_depth`)**: 10 – Controla el riesgo de sobreajuste.
  - **Mínimo de muestras por Hoja (`min_samples_leaf`)**: 4 – Evita reglas demasiado específicas.
  - **Mínimo de muestras para dividir (`min_samples_split`)**: 10 – Asegura divisiones significativas.
- **Logistic Regression:**
  - **Regularización (`C`)**: 100 – Menos regularización para ajustar más los datos.
  - **Solucionador (`solver`)**: `lbfgs` – Eficiente para tamaños moderados de datos.
- **Support Vector Machine (SVM):**
  - **Regularización (`C`)**: 10 – Permite un ajuste más flexible al conjunto de entrenamiento.
  - **Kernel**: `rbf` – Apto para capturar relaciones no lineales.
  - **Gamma**: `scale` – Ajusta la influencia de cada punto de entrenamiento.

4.2.2 Modelo de ensamble. El modelo de ensamble implementado es un clasificador de votación, denominado *Voting Classifier*. Este enfoque combina las predicciones de varios modelos de aprendizaje automático mediante un método de votación dura. En la votación dura, cada modelo individual dentro del ensamble vota por una clase, y la clase que recibe la mayoría de los votos es seleccionada como la predicción final del ensamble. Este método es efectivo para aumentar la robustez y precisión del modelo al reducir la varianza y los errores que podrían ser propios de un solo modelo predictivo. [ens [n. d.]]

El proceso de implementación del *Voting Classifier* implica varios pasos. Inicialmente, el ensamble se configura con una lista de los modelos, cada uno con su propio conjunto de hiperparámetros y entrenados independientemente. Posteriormente, este ensamble se entrena utilizando el conjunto de datos de entrenamiento, compuesto por pares de entradas y sus etiquetas correspondientes. Una vez entrenado, el clasificador de votación se utiliza para realizar predicciones sobre un conjunto de datos de prueba, y estas predicciones son evaluadas para determinar la precisión del modelo. Las métricas clave como la precisión, el recall y el puntaje F1 de cada clase, así como la precisión global, se calculan para proporcionar una evaluación completa del rendimiento del ensamble.

Al integrar diversos modelos, el *Voting Classifier* asegura un equilibrio entre diferentes estrategias de aprendizaje, resultando en una mejora general del rendimiento en la clasificación.

4.3 Evaluación

Ahora se explicarán los resultados de la evaluación de los modelos de clasificación utilizados. Esta sección desglosa el rendimiento de cada modelo individualmente, abordando métricas claves como la precisión, el recall, y el F1-score, así como el F2-score que nos da una perspectiva más centrada en la importancia del recall en nuestro análisis. Además se evaluó la efectividad del modelo de ensamble, que combina las los modelos individuales para mejorar la precisión y la robustez general de las predicciones.

Gaussian Naive Bayes

- Precisión General: 92.39%
- F2-score: 0.9242
- Comentarios: El modelo Gaussian Naive Bayes, a pesar de su simplicidad, mostró buena capacidad para clasificar correctamente las clases, con un equilibrio entre precisión y sensibilidad.

K-Nearest Neighbors (KNN)

- Precisión General: 93.105%
- F2-score: 0.9399
- Comentarios: KNN demostró ser muy efectivo, lo que se refleja en su alto F2-score.

Decision Tree

- Precisión General: 96.365%
- F2-score: 0.9721
- Comentarios: El árbol de decisión mostró el mejor rendimiento, teniendo la más alta precisión y F2-score entre todos los modelos evaluados.

Logistic Regression

- Precisión General: 93.45%
- F2-score: 0.9583

Support Vector Machine (SVM)

- Precisión General: 95.855%
- F2-score: 0.9676
- Comentarios: SVM fue altamente eficaz, con el segundo lugar de efectividad de todos los modelos.

Modelo de Ensamble: Voting Classifier

- Precisión General: 96.955%
- F2-score: 0.9695
- Comentarios: El modelo de ensamble superó a todos los modelos individuales, demostrando que la combinación de múltiples enfoques de modelado puede mejorar significativamente la precisión y la robustez general.

El F2-score es una variante del F1-score que se ajusta para ponderar el recall más que la precisión, específicamente haciendo que el recall sea dos veces más importante. Esta métrica es valiosa en contextos donde los falsos negativos, es decir, no detectar un caso positivo, tienen consecuencias más graves que los falsos positivos. En el ámbito del proyecto, esto significa tener más capacidad de no perder objetos celestes raros o de interés crítico, aunque esto implique aceptar un número mayor de falsos positivos. Por ejemplo, perder la detección de un nuevo quasar o de una galaxia distante debido a un modelo demasiado conservador podría significar perder información clave.

5 CONCLUSIONES

A lo largo de este proceso, se evaluaron varios modelos de aprendizaje automático utilizando una serie de métricas para determinar su eficacia en la clasificación de objetos celestes. Se optó por el F1-score como métrica principal para tener un equilibrio entre precisión y recall, lo cual es crucial en contextos donde es fundamental no solo capturar todos los casos positivos sino también mantener una alta precisión para evitar falsos positivos. Sin embargo, también fue importante la métrica F2-score, que pone un énfasis adicional en el recall, lo que es esencial en situaciones donde perder un descubrimiento podría tener consecuencias significativas. El usar estas métricas ayudó a identificar modelos que proporcionan un buen equilibrio y también evitar pasar por alto los casos positivos vitales.

El árbol de decisión y las Support Vector Machines (SVM) con kernel RBF demostraron ser muy eficaces. Estos modelos manejaron el desequilibrio de clases y generalizaron bien. La optimización de hiperparámetros desempeñó un papel crucial en maximizar el rendimiento de cada modelo. Mediante el uso de 'GridSearchCV', se exploraron distintos hiperparámetros, lo que resultó en mejoras significativas tanto en la precisión como en el recall. Este proceso es fundamental y muestra cómo es que los hiperparámetros pueden influir enormemente en la eficacia de los modelos.

Los aprendizajes obtenidos resaltan la importancia de comprender profundamente las métricas de evaluación, seleccionar los modelos adecuados, optimizar los hiperparámetros y la implementación de un método de ensamble.

REFERENCES

[n. d.]. 1.11. Ensembles: Gradient boosting, random forests, bagging, voting, stacking. https://scikit-learn.org/stable/modules/ensemble.html
[n. d.]a. Galaxies - NASA Science. https://science.nasa.gov/universe/galaxies/
[n. d.]. Image Gallery | SDSS. https://www.sdss4.org/science/image-gallery/
[n. d.]b. Stars - NASA Science. https://science.nasa.gov/universe/stars/
[n. d.]. What is 'red shift'? https://www.esa.int/Science_Exploration/Space_Science/What_is_red_shift
The Editors of Encyclopaedia Britannica. 2024. Electromagnetic spectrum | Definition, Diagram, Uses. https://www.britannica.com/science/electromagnetic-spectrum