Primera entrega de proyecto

POR:

Juan José Gomez Mejia

MATERIA:

Introducción a la inteligencia artificial

PROFESOR:

Raul Ramos Pollan

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

MEDELLÍN 2023

1. Planteamiento del problema

El descubrimiento de exoplanetas se ha convertido en una tarea muy importante y emocionante para la astronomía moderna; desde el descubrimiento del primer exoplaneta en 1995 por los astrónomos Michel Mayor y Didier Queloz, al que llamaron 51 Pegasi b, utilizando la técnica de velocidad radial, que es una técnica de detección indirecta que se basa en el análisis del efecto gravitacional de un planeta en su estrella anfitriona, realizando así la primera detección directa del espectro de luz visible reflejado por un exoplaneta; hasta la reciente detección espectral de transmisión atmosférica del exoplaneta WASP-39b capturado por el espectrógrafo de infrarrojo cercano del telescopio James Webb, revelando la primera evidencia clara de dióxido de carbono en la atmósfera de un planeta fuera del Sistema Solar.

Porque resulta que detectar exoplanetas, e incluso intentar obtener datos en espectro de luz visible es extremadamente difícil, debido a varios desafíos técnicos y astronómicos involucrados:

En primer lugar, los exoplanetas son objetos muy pequeños y débiles en comparación con sus estrellas anfitrionas. Debido a que los exoplanetas no emiten su propia luz, sino que reflejan la luz de su estrella, cualquier señal de un exoplaneta es enmascarada por el brillo abrumador de su estrella. Esto hace que sea difícil separar la señal del exoplaneta de la señal de la estrella, especialmente cuando el exoplaneta está muy cerca de su estrella anfitriona.

En segundo lugar, los exoplanetas están muy lejos, lo que hace que incluso los telescopios más potentes tengan dificultades para obtener datos detalladas de ellos. Para hacerlo, los telescopios deben ser capaces de detectar la luz de un exoplaneta que está separado por solo unos pocos píxeles de su estrella anfitriona, lo que requiere técnicas avanzadas de procesamiento de imágenes y una gran estabilidad en la observación.

Sin embargo, actualmente existe una grande lista de exoplanetas confirmados, siendo más de 5000, ubicados en aproximadamente más de 3650 sistemas planetarios, descubiertos entre las distintas misiones espaciales como Kepler, TESS, CHEOPS, y técnicas de detección como los tránsitos planetarios, velocidad radial, microlentes gravitatorias y técnicas de imagen directa.

2. Dataset

Los datos del proyecto vienen del sitio web Kaggle, con nombre **Kepler Exoplanet Search Results**, proporcionados por la NASA; son datos obtenidos por la misión espacial Kepler, capturados por el telescopio espacial Kepler lanzado en 2009, y operando hasta 2018, siendo la punta de la lanza en detección de exoplanetas, donde más de los cerca de 5000 candidatos planetarios encontrados hasta la fecha, más de 3.200 ahora han sido verificados, y 2.325 de estos fueron descubiertos por el Kepler.
De esta forma, el dataset se conforma de 50 columnas o características y 9564 filas o muestras y el objetivo principal del proyecto es poder predecir si un exoplaneta **detectado**, puede clasificarse en:

CANDIDATO: Se ha detectado un tránsito en los datos del telescopio, pero no se han realizado observaciones adicionales para confirmar si el tránsito es causado por un verdadero planeta.

CONFIRMADO: Se ha verificado mediante observaciones adicionales que el tránsito del candidato a exoplaneta es causado por un planeta real que orbita alrededor de una estrella.

FALSO POSITIVO: Se ha determinado que el tránsito del candidato a exoplaneta es causado por una señal falsa, como una estrella de fondo, un artefacto o una perturbación en los datos.

Teniendo así un problema de clasificación múltiple.

NOTA:
*(De acuerdo a la cantidad de muestras que son **FALSO POSITIVO** respecto a las demás, puede que se tenga que eliminar y trabajar sólo con* **CANDIDATO** y **CONFIRMADO***)*

Por otra parte, entre las características más importantes están:

- ***koi_disposition***: El estado de la KOI en función de la validación de su existencia. Puede ser "CANDIDATE", "CONFIRMED", o "FALSE POSITIVE". ***(La salida del dataset)***
- ***kepid***: El ID de Kepler para el objetivo observado.
- ***kepoi_name***: El nombre del planeta Kepler Object of Interest (KOI) asignado por el equipo de Kepler.
- ***koi_pdisposition:*** La probabilidad de que el planeta tenga una disposición positiva (confirmado) según un modelo estadístico. ***(Otra posible salida del dataset)***

NOTA:
Como se tienen 2 salidas, se excluirá esta columna y se trabajará con ***koi_disposition****. A fin de cuentas,* ***koi_pdisposition*** *sólo es una columna de decisión provisional sobre la existencia de exoplanetas.*

- ***koi_score***: La confiabilidad de la detección de un tránsito planetario en un sistema estelar.
- ***koi_period***: El período orbital del planeta, en días.
- ***koi_time0bk***: El tiempo de tránsito del primer tránsito, en días julianos.
- ***koi_duration***: La duración del tránsito, en horas.
- ***koi_prad***: El radio del planeta, en radios terrestres.
- ***koi_teq***: La temperatura de equilibrio del planeta, en grados Kelvin.
- ***koi_insol***: La insolación recibida por el planeta, en unidades de la Tierra.
- ***koi_steff***: La temperatura efectiva de la estrella, en Kelvin.
- ***koi_slogg***: La gravedad superficial de la estrella, en cm/s^2.
- ***koi_srad***: El radio de la estrella, en radios solares.
- ***koi_kepmag***: La magnitud aparente en banda Kepler del objetivo observado.
- ***koi_depth***: La profundidad del tránsito, en partes por millón (ppm).
- ***koi_count:*** El número de planetas en el sistema estelar.
- ***koi_period_err1:*** El error positivo en el período orbital del planeta, en días.
- ***koi_period_err2***: El error negativo en el período orbital del planeta, en días.
- ***koi_impact***: Parámetro de impacto estelar, que mide la distancia mínima entre el centro de la estrella y el centro del planeta en el momento del tránsito, en unidades de radio estelar.
- ***koi_smet:*** Metalicidad estelar, que mide la cantidad de elementos más pesados que el helio, presentes en la estrella anfitriona, en unidades de logaritmo de la relación con el Sol.
- ***koi_srho:*** Densidad estelar, que mide la densidad de la estrella anfitriona, en unidades de g/cm^3.
- ***koi_tce_plnt_num***: Número de planeta en el Sistema Multiplanet Transiting Candidate (MTC) al que pertenece la KOI.
- ***koi_quarters***: El trimestre en que se observó el objetivo.
- ***koi_disposition_score***: La puntuación de confianza asignada a la KOI por el equipo de validación de Kepler.
- ***koi_fpflag_nt:*** El número de veces que la KOI cruzó el borde de la máscara de destino.
- ***koi_fpflag_ss***: La cantidad de detecciones estadísticamente significativas de tránsito secundario.

3. Métricas

Para clasificar exoplanetas detectados como candidatos, confirmados o falsos positivos, es importante seleccionar las métricas adecuadas para evaluar el rendimiento del modelo.

En este caso, como se trataría de un problema de clasificación, se pueden utilizar métricas como la AUC, el Recall o la Precisión.

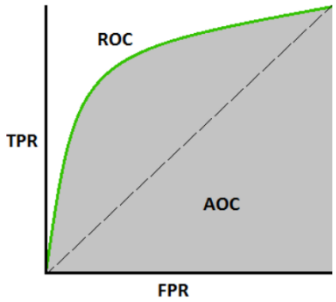
La métrica AUC (Área Bajo la Curva) se utiliza para evaluar la capacidad del modelo para distinguir entre dos clases: positiva y negativa. Un valor de AUC cercano a 0,5 indica un modelo que clasifica al azar, mientras que un valor cercano a 1 indica un modelo muy preciso. En general, un modelo con AUC mayor que 0,8 se considera que tiene un buen desempeño.

La fórmula para el cálculo del AUC es la siguiente:

$$AUC = \int TPR \, d(FPR)$$

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$



Donde TPR es la tasa de verdaderos positivos y FPR es la tasa de falsos positivos. La integral se realiza sobre el rango completo de FPR. La curva ROC se utiliza para obtener los valores de TPR y FPR para diferentes umbrales de clasificación, y se traza en un gráfico con TPR en el eje y y FPR en el eje x. El AUC se calcula a partir de esta curva, y proporciona una medida de la capacidad del modelo para distinguir entre las dos clases; por lo tanto, en nuestro caso, un valor alto de AUC indicaría que el modelo es capaz de distinguir entre las dos clases con alta precisión.

Por otra parte, **la métrica Recall** (o tasa de verdaderos positivos, TPR) es la métrica de evaluación también utilizada en clasificación que mide la capacidad del modelo para identificar correctamente las instancias positivas en relación al total de instancias positivas en el conjunto de datos.

La fórmula para calcular el recall es la siguiente:

$$Recall = \frac{TP}{TP + FN}$$

Donde TP (True Positives) es el número de instancias positivas que fueron correctamente clasificadas como positivas por el modelo, y FN (False Negatives) es el número de instancias positivas que fueron incorrectamente clasificadas como negativas por el modelo. Así, en general, un alto valor de recall indica que el modelo está identificando correctamente la mayoría de las instancias positivas, mientras que un bajo valor de recall indica que el modelo está perdiendo muchas instancias positivas, o que tiene una alta tasa de falsos negativos; en nuestro caso, el recall mediría la proporción de exoplanetas confirmados que se predicen correctamente.

Y finalmente se podría utilizar **la métrica de Precisión**, porque un alto grado de precisión es deseable; ya que la identificación de nuevos exoplanetas es un desafío complejo y se necesita una alta veracidad para garantizar que los resultados sean confiables. En general, un porcentaje de Precision superior al 90% sería deseable para este tipo de problema de clasificación. Así, deseamos que de los exoplanetas detectados más del 90% sean identificados correctamente como tal.

La fórmula para calcular la precisión es la siguiente:

$$Precision = \frac{TP}{TP + FP}$$

Donde TP (True Positives) es el número de predicciones positivas que fueron correctamente clasificadas como positivas por el modelo, y FP (False Positives) es el número de predicciones positivas que fueron incorrectamente clasificadas como positivas por el modelo.

4. Desempeño

En retrospectiva, si resulta factible utilizar modelos de machine learning para la detección o confirmación de exoplanetas; porque podremos mejorar las técnicas de identificación de patrones y características en los datos que no son evidentes para los científicos a simple vista, como las técnicas de detección de e identificación de nuevos exoplanetas que de otra manera podrían haber pasado desapercibidos.

Además, un modelo de machine learning puede utilizarse como una herramienta de selección de objetivos para futuras misiones espaciales. El modelo puede ser entrenado en datos de exoplanetas conocidos para identificar características que sugieran la presencia de un exoplaneta similar a la naturaleza de alguno del sistema solar, o incluso la tierra. A partir de ahí, se pueden identificar posibles candidatos para misiones futuras, dirigir los recursos de manera más efectiva y en última instancia, mejorar nuestras capacidades para explorar el universo.

5. Bibliografía

- Kepler Objects of Interest. https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=koi

- Nasa. “Kepler Exoplanet Search Results.” Kaggle, 10 Oct. 2017. www.kaggle.com/nasa/kepler-exoplanet-search-results.

- “Overview.” NASA, NASA, 2 Apr. 2021. https://exoplanets.nasa.gov/what-is-an-exoplanet/planet-types/overview/

- Malik, A. P. Moster, B. Obermeier C “Exoplanet Detection using Machine Learning.” https://arxiv.org/pdf/2011.14135.pdf

- Clayton, G. Manry, B. Rafiqi, S “Machine Learning Pipeline for Exoplanet Classification” 2019 https://scholar.smu.edu/cgi/viewcontent.cgi?article=1070&context=datasciencereview