Proyecto 1 Astrofísica Computacional: Predicción de eclipses

Juan Sebastián Martínez Arévalo[1], Emmanuel Arias Polanco[1], y Gabriel Suarez Cadena[1]

[1]Departamento de Física, Universidad Nacional de Colombia, Bogotá, Colombia

Octubre 2023

1. Introducción

Durante toda nuestra vida, y desde los inicios mismos de nuestra historia, convivimos con este interesante y notorio fenómeno que, durante mucho tiempo, dio lugar a curiosas e interesantes historias que intentaban explicar su funcionamiento. Hoy, entendemos plenamente su funcionamiento, y, a forma de corolario de la comprensión de la mecánica celeste, somos capaces de predecirlo de forma precisa para nuestro ocio o para usarlo en investigación científica, como se hizo en 1919 para someter a prueba por primera vez la Relatividad General de Einstein.

Es de considerable importancia la creación de un algoritmo que permita determinar la ocurrencia de un eclipse, bien sea lunar o solar, en un tiempo futuro, como también lo es la implementación de problemas astrofísicos y su solución mediante el uso de modelos de inteligencia artificial; por esto el principal objetivo del proyecto es la implementación de un algoritmo que, haciendo uso de sistemas de inteligencia artificial, pueda determinar la ocurrencia de eclipses futuros y medir la precisión con que predice dichos sucesos. A pesar de la fuerte comprensión teórica que se tiene actualmente sobre ese y otros fenómenos de la mecánica celeste, resulta difícil a nivel de cálculos y poder de cómputo, desarrollar algoritmos eficientes basados en nuestras teorías que sean capaces de, al mismo tiempo, obtener predicciones precisas y obtener resultados en tiempos de cómputo racionales. Por otro lado, con el actual desarrollo en el campo de la computación y el Aprendizaje Automático (Inteligencia Artificial), se ha notado que, bajo este paradigma, es posible entrenar modelos de aprendizaje que replican de forma correcta, y en tiempos de cálculo mucho más razonables, teorías y leyes físicas de nuestro universo, resultando en modelos de gran interés cuando se quiere replicar la realidad física de forma rápida (como se requiere por ejemplo en el campo de gráficos por computadora)

Es así, como M. Sferdian, M. Frincu (2021) Mara Sferdian1, 2021, muestran la mejora computacional obtenida entre el métodos clásico implementado en un algoritmo de fuerza

bruta, y el entrenamiento de un modelo de Machine Learning para la predicción de eclipses. En el trabajo actual, se replica el trabajo hecho en el artículo para entrenar el modelo de aprendizaje automático.

2. Datos

La página web de la Administración Nacional de Aeronáutica y el Espacio de los Estados Unidos (NASA, por sus siglas en inglés) dispone para el dominio público de un catálogo de eclipses solares y lunares que abarcan un lapso de tiempo de alrededor de cinco mil años, de tal forma que para cada eclipse es posible establecer la fecha de su apariencia y un factor geométrico de separación. La base de datos a usar contiene, en un rango de fechas a establecer, todos los días, un valor de separación desde el punto de vista de un observador conveniente y si ese día hubo eclipse o no.

En el trabajo desarollado por nosotros, implementaremos solo los métodos basados en Machine Learning descritos en el artículo original, a ser, K-Nearest Neighbors, Logistic Regression para clasificación, y Random Forest, además un modelo basado en Redes Neuronales para comparar los resultados. De lo anterior, se obtuvo el mejor resultado para el modelo -simple- basado en redes neuronales, y para este modelo se realizó predicciones de eclipses dada una fecha de entrada como cadena de caracteres siguiendo el formato YYYY-MM-DD.

El método descrito en el artículo para la identificación de eclipses (pag. 8, M. Sferdian, M. Frincy (2021)), se basa en los métodos VSOP87 y ELP-2000/82 para la determinación de coordenadas solares y lunares en el cielo. A partir de ellas, se determina, para cada día en el conjunto de datos original, y para distintas locaciones en la esfera terrrestre cada día, la separación mínima entre los dos cuerpos celestes, conociendo sus radios y teniendo en cuenta las condiciones de un eclipse solar y uno lunar (mes sideral y mes dracónico). Este cálculo representa el parámetro de identificación de eclipses basado en machine learning.

3. Métodos y algoritmos

Basándonos en el paper de Sferdinan, M, et. al. y debido a que se busca establecer si, bajo ciertas condiciones, un suceso (eclipse -bien sea solar o lunar-) ocurrió, es pertinente establecer modelos de aprendizaje de clasificación binaria. De esta forma, se plantea la implementación de modelos de regresión logística, árboles de decisión, random forest y keras, para evaluar y comparar su posible desempeño en la clasificación de eclipses basándose en la separación y la fecha.

4. Implementacion

Primero haremos la lectura de datos y la definición de los conjuntos de entrenamiento y de prueba. El conjunto de datos original contiene información desde el año 1999 a.C, hasta el año 2100 d.C, estos datos son obtenidos de bases de datos de la NASA (https://eclipse.gsfc.nasa.gov/), que comprenden un de eclipses solares y lunares desde el año 1999 a.C hasta el año 3000 d.C. Para el entrenamiento de nuestros modelos, usaremos

un subconjunto del conjunto de datos descargado, que es, a su vez, un subconjunto del los datos originales de la NASA; estos conjuntos de datos se dividen como se describe a continuación.

4.1. Datos de entrenamiento

Consiste en datos de eclipses solares desde el primero de enero de 1950 (1950-01-01), hasta datos del treinta y uno de diciembre de 2019 (2019-12-31). En estos datos se puede corroborar la congruencia con los datos observacionales, como el pasado eclipse del 14 de octubre del 2023.

4.2. Datos de prueba

l conjunto de prueba consiste el resto de datos de eclipses solares, desde el primero de enero del 2020 hasta el 31 de diciembre del 2100. Aquí, igualmente, se puede corroborar la congruencia de los datos al comparar con, por ejemplo, los futuros eclipses para el 8 de abril de 2024 y el 29 de marzo de 2025.

Con estos conjuntos de prueba, implementaremos los distintos modelos de Machine Learning utilizando las implementaciones en la librería scikit-learn. Los resultados obtenidos se discuten luego de las implementaciones realizadas.

5. Análisis

5.1. Clasificación KN

El agoritmo K-NN es un algoritmo no supervisado de agrupación de datos, y será la primera implementación que haremos. Los parámetros establecidos antes del entrenamiento del modelo, son resultado de prueba y error hasta obtener una buena precisión. Algunos datos son dejados como el valor por defecto implementado en la librería.

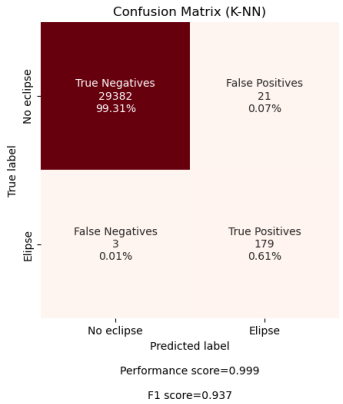

Figura 1: Matriz de confusión método de clasificación KNN, se observa una precisión d e predicción del 99.2 %

5.2. Regresión Logística

La regresión logística resulta ser, en este caso, una aproximación más natural al problema, dado el sistema de clasificación binaria al que apuntamos. En esta implementación todos los parámetros se dejaron iguales a los parámetros por defecto en la librería.

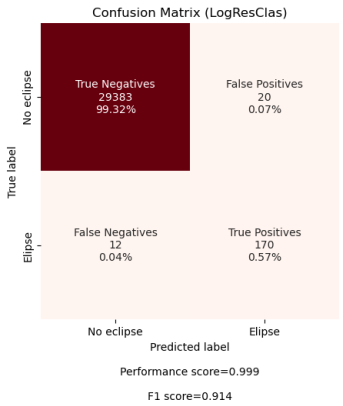

Figura 2: Matriz de decisión regresión logística

5.3. Arboles de decisión

De la misma forma que para la clasificación logística, el algoritmo de bosques aleatorios resulta una aproximación más antural al problema, así como a su vez, una mejora en el algoritmo de regresión logística, dados los criterios de comparación empleados. En este caso, parámetros como el número de estimadores, no mostraron mayor relevancia a la hora de ejecutar la clasificación.

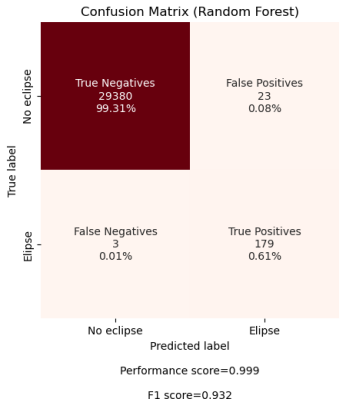

Figura 3: Matriz de Confusión para el algoritmo Árbol de decisión,observa una presición de 89 %

5.4. Red Neuronal Multicapa

Finalmente, hacemos uso de un algoritmo de clasificación basado en redes neuronales con 5 capas ocultas de 5, 8 y 8 neuronas respectivamente, haciendo uso de la función de activación relu y el algoritmo de optimización adam. Este número de capas y de neuronas son escogidas a través de prueba y error que ofrece un modelo eficiente. Dada la naturaleza de los algoritmos basados en redes neuronal, podemos esperar que esta optimización sea buena, pero no podemos, apriori, estimar qué tanto respecto a los otros métodos.

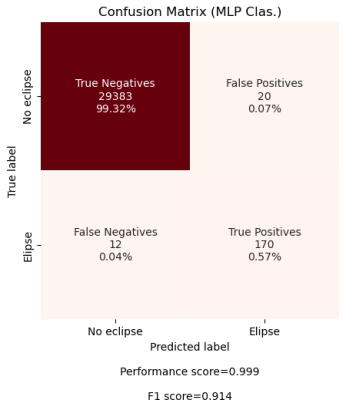

Figura 4: Caption

6. Resultados

De todos los métodos implementados, el método basado en redes neuronales predijo correctamente un total (TP + TN) de 29553 datos de eclipses, que corresponden a aproximadamente el 99,89 % del total de datos. Además de esto, se obtuvo el mejor rendimiento, en términos de eficiencia en tiempos de cálculo, para el algoritmo de clasificación logística, lo cual es esperable si consideramos la naturaleza binaria del problema de clasificación en el que nos encontramos; de la misma forma, y como era de esperarse, el modelo basado en bosques aleatorios, representa un algoritmo con un rendimiento inferior la regresión logística, pero con una mejor precisión, prediciendo correctamente el 99,92 % de los datos (TP + TN = 29559), respecto al 99,89 % (TP + TN = 29553). El modelo basado en K-NN resultó tener la precisión y rendimientos más bajos, probablemente debido a que el problema no está íntimamente relacionado a su campo de aplicación, a pesar de esto, este algoritmo resulta tener menos equivocaciones al momento de identificar falsos positivos y falsos negativos, obteniendo, en cada caso FP = 21, FN = 3.

El problema con los falsos negativos y falsos positivos puede deberse a la forma en la que se determina si hay un eclipse solar o no. Podemos notar que la mayoría de errores de predicción se encuentran a la hora de identificar aproximaciones sol-luna que no resultan ser eclipses. Es importante ver también que el mejor rendimiento se tiene a la hora de determinar cuándo no hubo un eclipse, y esto presenta un sesgo, dada la cantidad mínima de eclipses respecto al total de días analizados; sin embargo, el porcentaje de identificación de eclipses tampoco

está representada notablemente por los porcentajes de errores. Si tenemos en cuenta este sesgo, y el porcentaje de desaciertos, el mejor resultado se obtiene para el modelo basado en K-NN, y esto se puede deber a que este modelo es capaz de encontrar nuevas características entre los conjuntos de datos que no son contempladas en el modelo original de identificación de eclipses.

Referencia

[1] Sferdian, M., Frincu, M. (2021). When old meets new: evaluating numerical and Machine Learning based eclipse prediction methods. Romanian Astronomical Journal, 31(2), 133-152.

[2] https://github.com/CosmeticMichu/Eclipse-Prediction-by-ML/blob/master/SolarEclipses.ipynb