UNIVERSIDAD TECNOLÓGICA DE LA MIXTECA

Clasificación de asteroides potencialmente peligrosos mediante una máquina de soporte vectorial

TESIS
PARA OBTENER EL TÍTULO DE:

LICENCIADO EN MATEMÁTICAS APLICADAS

PRESENTA:

OSCAR RAMOS ORTIZ

DIRECTOR DE TESIS:

DR. TOMÁS PÉREZ BECERRA

HUAJUAPAN DE LEÓN, OAXACA.
SEPTIEMBRE DE 2024

Índice general

Índice de figuras

Lista de tablas

Agradecimientos

Quiero expresar mi agradecimiento a todas las personas que me brindaron su apoyo a lo largo de mi carrera, en especial mi más sincero agradecimiento a mi familia por su dedicación a mi educación, de no haber contado con su respaldo nada de esto habría sido posible.

Del mismo modo, extiendo mi gratitud a cada uno de mis profesores, quienes me instruyeron y estuvieron dispuestos en todo momento a brindarme orientación, así como a mi director de tesis el Dr. Tomás Pérez Becerra, cuya orientación y compromiso fueron esenciales para alcanzar este logro.

Introducción

En un universo vasto y en constante movimiento, los asteroides representan una fascinante y a veces preocupante parte del paisaje cósmico. Estos cuerpos rocosos, algunos de los cuales son lo suficientemente grandes como para causar daños significativos en caso de impacto con la Tierra, han despertado el interés y la atención de científicos y entusiastas del espacio por igual. Entre ellos, un grupo específico de asteroides, denominados asteroides potencialmente peligrosos (PHA, por sus siglas en inglés), ha sido objeto de un escrutinio particular debido a su capacidad para representar una amenaza real para nuestro planeta.

En este contexto, el presente proyecto se enfoca en abordar el desafío de identificar y clasificar asteroides potencialmente peligrosos mediante el uso de técnicas avanzadas de aprendizaje automático. Específicamente, se centrará en el empleo de una máquina de soporte vectorial (SVM, por sus siglas en inglés), una técnica de clasificación ampliamente utilizada y conocida por su eficacia en la separación de clases en conjuntos de datos complejos.

El objetivo principal de este proyecto es desarrollar un modelo predictivo que pueda distinguir con precisión entre asteroides potencialmente peligrosos y asteroides que no representan una amenaza significativa para la Tierra. Para lograr esto, se aprovecharán una variedad de características y parámetros orbitales de asteroides recopilados a través de observaciones astronómicas y datos de seguimiento.

En el desarrollo de la tesis, se describirán en detalle el contexto del problema, los datos utilizados, la metodología propuesta, los resultados esperados y las implicaciones potenciales de este proyecto. Específicamente, la tesis se divide en los siguientes capítulos:

Capítulo 1 Se realiza un análisis exploratorio y visualización de datos. En este capítulo se encuentran una serie de diagramas que permiten extraer algunas conclusiones parciales, entre ellas, que se requiere hacer una imputación a los datos faltantes, esta es la motivación del capítulo siguiente.

Capítulo 2 Al existir diversos valores faltantes en algunas variables que se consideran relevantes, se realiza un proceso de imputación de datos mediante el algoritmo EM, con lo que se logra tener una base completa. Cabe resaltar que el algoritmo EM es descrito a detalle.

Capítulo 3 Para determinar las variables relevantes que afectan a la variable elegida como objetivo (específicamente, si un asteroide es potencialmente peligroso o no lo es), en este capítulo se realiza una selección de características.

Capítulo 4 Finalmente, el modelo de máquina de soporte vectorial es entrenado con la base completa y se utilizan tres distintos kernel para el modelo. En este mismo capítulo se

encuentra la validación de cada uno de ellos y se determina el de mejor desempeño.

A manera de justificación de la investigación, la clasificación de asteroides PHA es un campo de investigación crítico en el ámbito de la astronomía y la protección planetaria. La capacidad de identificar y monitorear asteroides potencialmente peligrosos no solo contribuye a nuestra comprensión del sistema solar, sino que también nos proporciona información valiosa para la planificación y mitigación de posibles impactos astronómicos. A través de este esfuerzo, se busca prevenir y gestionar los riesgos asociados con asteroides potencialmente peligrosos, contribuyendo así a la seguridad y la comprensión del cosmos mediante la modelación matemática aplicada en la ciencia de datos.

Capítulo 1

Análisis exploratorio y visualización de los datos

El análisis exploratorio de datos (o EDA por sus siglas en inglés) se utiliza principalmente para ver qué pueden revelar los datos más allá de las tareas de modelado formal o prueba de hipótesis, y para proporcionar una mejor comprensión de las variables en un conjunto de datos y las relaciones entre ellas. También puede ayudar a determinar si los métodos estadísticos que se están considerando para el análisis de datos son apropiados, esto también ayuda a identificar errores que son obvios. La tecnología EDA fue desarrollada originalmente por el matemático estadounidense John Tukey en la década de 1970 y sigue siendo una técnica ampliamente utilizada en el proceso de descubrimiento de conocimiento.

1.1. Descripción del conjunto de datos

La base de datos que se analizará es oficialmente mantenida por el *Jet Propulsion Laboratory* del *California Institute of Technology* de la *National Aeronautics and Space Administration* (NASA). Esta base incluye datos relacionados con asteroides, es pública y está disponible en el sitio `https://ssd.jpl.nasa.gov/sbdb_query.cgi`
Esta base es denotada en esta investigación como D, cuenta con un total de 43 características de asteroides, a continuación se muestran algunas de ellas:

- SPK-ID: Identificador del cuerpo celeste.
- Object ID: Identificador interno de la base de datos.
- Object fullname: Nombre completo del objeto.
- pdes: Designación primaria del objeto.
- name: Nombre del objeto en el estilo de la *International Astronomic Union*.
- neo: Etiqueta de objeto cercano a la tierra (*Near-Earth Object*).
- *pha*: Etiqueta de asteroide potencialmente peligroso (*Potentially Hazardous Asteroid*).
- H: Magnitud absoluta.

- Diameter: Diámetro del asteroide (de la esfera equivalente) en kilómetros.
- Albedo: Albedo geométrico (porcentaje de radiación que cualquier superficie refleja respecto a la radiación que incide sobre ella perpendicular a la superficie).
- Diameter_sigma: incertidumbre 1-sigma en el diámetro del objeto en kilómetros.
- Orbit_id: Identificador de la órbita.
- Epoch: Época de osculación en forma de día juliano modificado.
- Equinox: Equinoccio desde el marco de referencia (Cada uno de los dos puntos de la esfera celeste en que el ecuador corta la eclíptica).
- e: Excentricidad.
- a: Semi-eje mayor en unidades astronómicas [1].
- q: Distancia del perihelio en unidades astronómicas.
- i: ángulo de inclinación con respecto al plano de la eclíptica x-y.
- tp: Tiempo de paso del perihelio en tera billones de años.
- moid_ld: Distancia mínima de intersección de la órbita de la Tierra en unidades astronómicas.
- **epoch**: (época en español) se refiere a un punto específico en el tiempo que se utiliza como referencia para describir la posición o el estado de un objeto celeste en su órbita o movimiento aparente en el cielo. Esencialmente, la época proporciona un marco de referencia temporal para las observaciones astronómicas y los cálculos relacionados con el movimiento de objetos celestes.

Formalmente, la base de datos D es de la forma que muestra la definición 1.1 extraída de [1].

Definición 1.1. *(Datos multidimensionales) Un conjunto de datos multidimensionales D es un conjunto de n registros, $\overline{X_1}, \ldots, \overline{X_n}$, tal que cada registro $\overline{X_i}$ contiene un conjunto de d características denotadas por $(x_i^1, \ldots, x_i^d)$.*

Para la base de datos utilizada D se tiene un total de $n = 958,523$ registros más una fila de nombres de las características, esto es, es el número de filas es n, y $d = 43$, es el número de columnas. Específicamente,

$$\overline{X}_1 = (a0000001, 2000001, 1Ceres, \ldots, MBA, 0.43301),$$

$$\vdots$$

$$\overline{X}_{958,523} = (bT3S2678, 3246553, (2678T-3), \cdots, MBA, 0.26980).$$

Nótese que nuestra base se conforma de datos no orientados a la dependencia, datos textuales, enteros y flotantes, esto es, es una base de datos mixta al incluir datos categóricos y numéricos.

[1] unidad astronómica de distancia (au) 1 au = 149 597 870 700 metros.

1.2. Clases de Asteroides

Los asteroides dentro de la base cuentan con una clasificación, las cuales son "MBA", "OMB", "MCA", "AMO", "IMB", "TJN", "APO", "ATE", "CEN", "AST" y "TNO", las cuales son descritas a detalle en la siguiente lista.

1. MBA (*Main-belt Asteroid*): Asteroides del cinturón principal. Asteroides con elementos orbitales restringidos por $2.0\ au < a < 3.2\ au$ y $q > 1.666\ au$.
2. OMB: Asteroide del cinturón principal exterior con elementos orbitales restringidos por $3.2\ au < a < 4.6\ au$.
3. MCA: Asteroides que cruzan la órbita de Marte con $1.3\ au < q < 1.666\ au$ y $a < 3.2\ au$.
4. AMO (Amor): Órbitas de asteroides cercanas a la Tierra similares a las del asteroide 1221 Amor, es decir, $a > 1.0\ au$ y $1.017\ au < q < 1.3\ au$.
5. IMB: Asteroide del cinturón principal interno con elementos orbitales dentro del intervalo $a < 2.0\ au$ y $q > 1.666\ au$.
6. TJN (Troyano de Júpiter): Son asteroides atrapados en los puntos de Lagrange[2] L4 y L5 de Júpiter con valores orbitales $4.6\ au < a < 5.5\ au$ y $e < 0.3$.
7. APO (Apolo): Órbitas de asteroides cercanas a la Tierra que cruzan la órbita terrestre similar a la del asteroide Apolo 1862 con valores $a > 1.0\ au$ y $q < 1.017\ au$.
8. ATE (Aten): Asteroide cercano a la Tierra orbita similar a la de 2062 Aten, cuyos valores orbitales se restringen a $a < 1.0\ au$ y $q > 0.983\ au$.
9. CEN (Centauro): Objetos con órbitas entre Júpiter y Neptuno $5.5\ au < a < 30.1\ au$.
10. AST (Asteroide): La órbita del asteroide no coincide con ninguna clase de órbita definida.
11. TNO (Objeto transneptuniano): Objeto del sistema solar cuya órbita se ubica parcial o totalmente más allá de la órbita del planeta Neptuno, se caracterizan por tener el valor $a > 30.1\ au$.

1.3. El problema de clasificación

El problema de clasificación en este estudio se puede plantear como determinar si un asteroide es potencialmente peligroso o no, esta variable binaria será la etiqueta de clase y se encuentra dentro del conjunto de datos en la columna llamada *pha*, que contiene los valores Y para los potencialmente peligrosos y N para los que no lo son, por lo que será esta la característica

[2]De acuerdo con [6], *"normalmente uno de los cuerpos será mucho mayor que el otro, como en el caso del Sol y un planeta, de modo que el centro de masas (o centro de gravedad) casi coincide con el objeto más masivo. Si la órbita del objeto menor es circular, entonces hay cinco puntos concretos, los puntos de Lagrange, en los que sería posible colocar un tercer objeto de manera que quede en equilibrio, en una posición estacionaria vista desde los otros dos."* Véase la Figura 1.1.

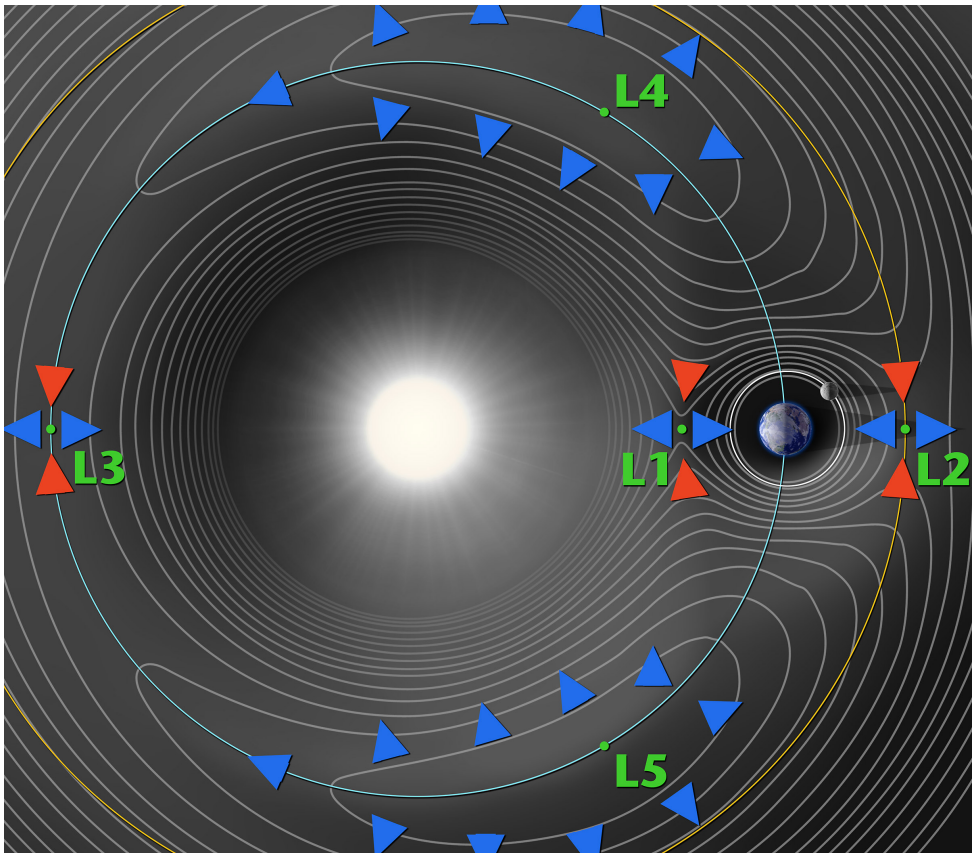


Figura 1.1: Puntos de Lagrange para el sistema Tierra-Sol (imagen obtenida de [3]).

objetivo. Por lo anterior, el problema se aborda como de aprendizaje supervisado, donde las relaciones de las características restantes en los datos respecto a esta característica especial son aprendidos.

Los datos utilizados para aprender estas relaciones se les llama *datos de entrenamiento*. Luego de entrenado, el modelo de aprendizaje será utilizado para determinar las etiquetas de clase para los registros que se introduzcan sin esta etiqueta, esto es, al introducirle nuevos registros donde se desconoce si pertenecen a un asteroide potencialmente peligroso o no, el programa debe estimar la etiqueta Y o N para cada caso, es decir, deberá agregarlos a una clasificación.

El problema de clasificación se plantea formalmente como sigue:

Definición 1.2. *(Clasificación de datos) Dadas una $m \times d$ matriz de datos de entrenamiento D_T, la cual es un subconjunto propio de la matriz D, esto es, $m < n$, y una etiqueta de clase en el conjunto $\{N, Y\}$, el cual, lo consideraremos como $\{0, 1\}$ mediante la transformación $N \to 0$ y $Y \to 1$, el problema es crear un modelo de entrenamiento $\mathcal{M}$, el cual será utilizado para predecir la etiqueta de clase de un registro d-dimensional $\overline{Y} \notin D$.*

1.4. Análisis de correlación

El análisis de correlación es una técnica estadística que se utiliza para evaluar la fuerza y la dirección de la relación lineal entre dos variables cuantitativas. En otras palabras, el análisis de correlación examina si existe una asociación sistemática entre dos variables, de modo que

cuando una variable cambia, la otra también tiende a cambiar de manera consistente.

La medida más común utilizada en el análisis de correlación es el coeficiente de correlación, que cuantifica la fuerza y la dirección de la relación entre las variables. El coeficiente de correlación más conocido es el de Pearson, que varía entre -1 y 1. Un valor de 1 indica una correlación perfecta positiva (ambas variables aumentan juntas), un valor de -1 indica una correlación perfecta negativa (una variable aumenta mientras la otra disminuye), y un valor de 0 indica ausencia de correlación lineal.

En este caso, se realizó un análisis de correlación entre parejas de variables con el objetivo de determinar la relación entre la variable objetivo *pha* con el resto y también determinar cuales variables tienen correlación exacta, esto debido a que si dos variables están altamente correlacionadas, tendrán el mismo efecto dentro del estudio.

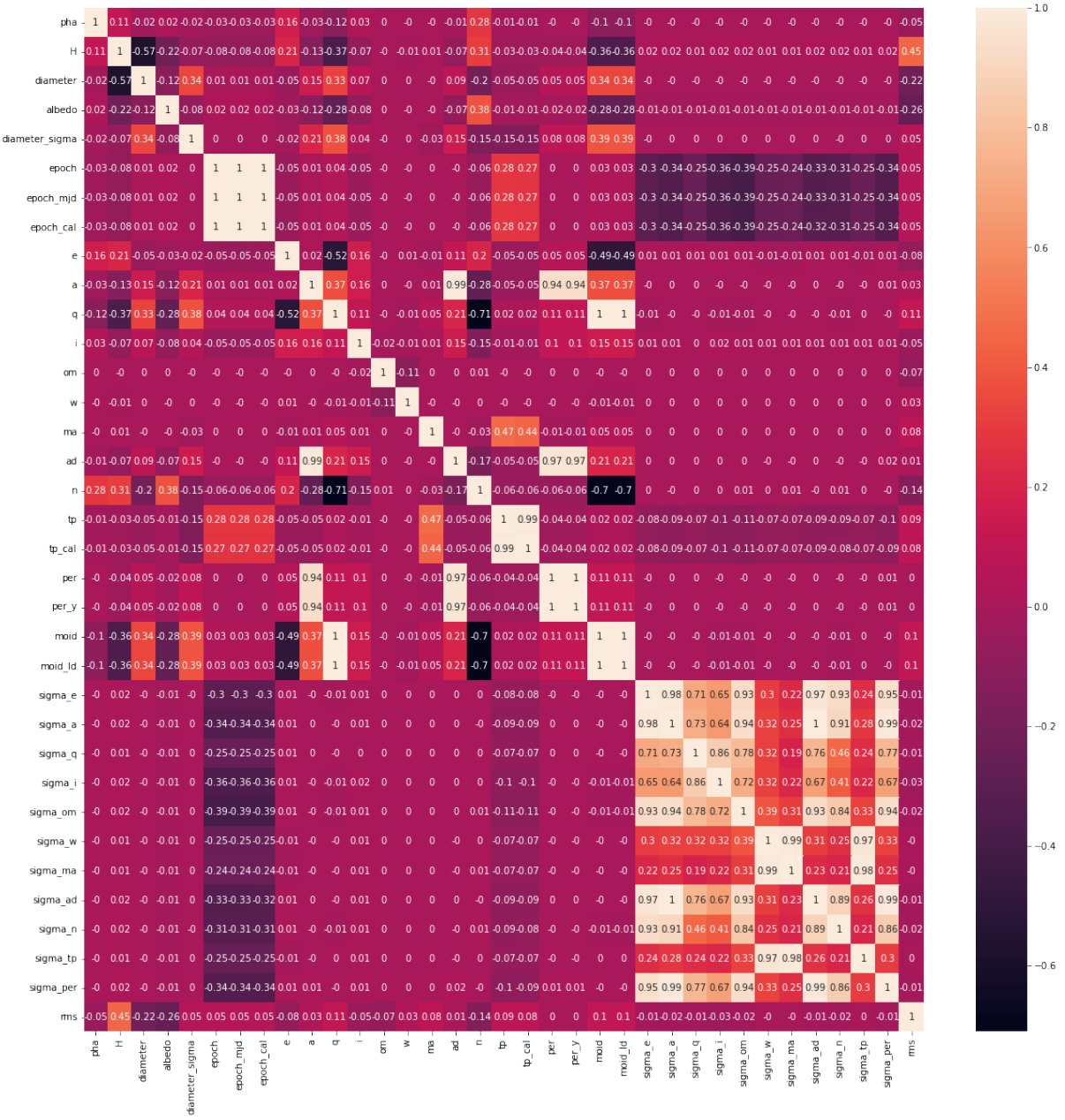


Figura 1.2: Matriz de correlación de la base de datos D.

En la Figura 1.2 se muestra el resultado del análisis de correlación, se observa que la variable *pha* no está correlacionada con el resto de las variables, por lo que este método no es concluyente. De aquí la necesidad de aplicar un modelo de aprendizaje automático que determine la relación de la variable objetivo con el resto.

1.5. Visualización de datos

La base se conforma de d distintos tipos de datos (características), que son las columnas de la matriz D, en esta sección se estarán analizando las más destacadas para la investigación. Un proceso relevante dentro del análisis exploratorio de cualquier base de datos es verificar el

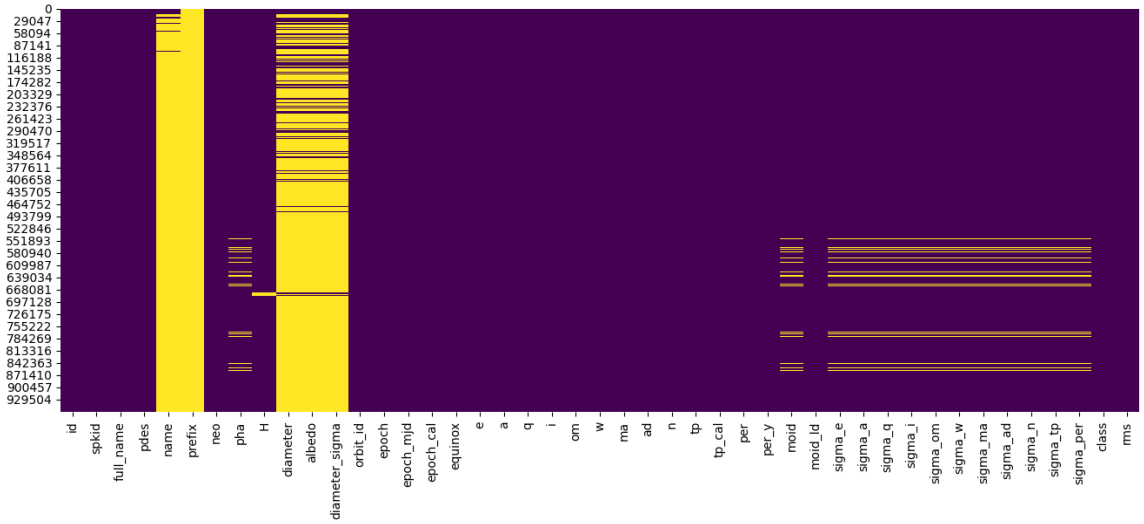


Figura 1.3: Mapa de calor de la base de datos D. El color amarillo representa los datos faltantes y el morado los datos observados.

porcentaje de datos faltantes en cada columna. Obsérvese la Figura 1.3, donde se muestra un mapa de calor conformado de dos colores, rojo representa datos faltantes mientras que el color azul representa datos existentes. A simple vista, resaltan algunas columnas que gran parte de ellas son vacías, estas columnas son:

- name: Nombre del objeto en el estilo de la International Astronomic Union.
- prefix: Prefijo del nombre del asteroide.
- diameter: Diámetro del asteroide (de la esfera equivalente) en kilómetros.
- albedo: Albedo geométrico (porcentaje de radiación que cualquier superficie refleja respecto a la radiación que incide sobre ella perpendicular a la superficie).
- diameter_sigma: incertidumbre 1-sigma en el diámetro del objeto en kilómetros.

En las secciones posteriores se realizará la limpieza e imputación de datos, donde se retomarán estas características.
Algunas de las variables relacionadas con las órbitas son descritas a detalle en lo consecuente.

- **Class.** En la sección 1.2 se mostraron las clasificaciones de los asteroides en relación a su tamaño. En específico, respecto al valor del semi-eje mayor (a) y al valor de la distancia del perihelio (q). Es relevante en la investigación considerar esta característica debido a que algunos autores asignan una mayor peligrosidad a aquellos asteroides con mayor tamaño. En la Figura 1.4 se observa que aproximadamente el 90 % de los asteroides pertenecen

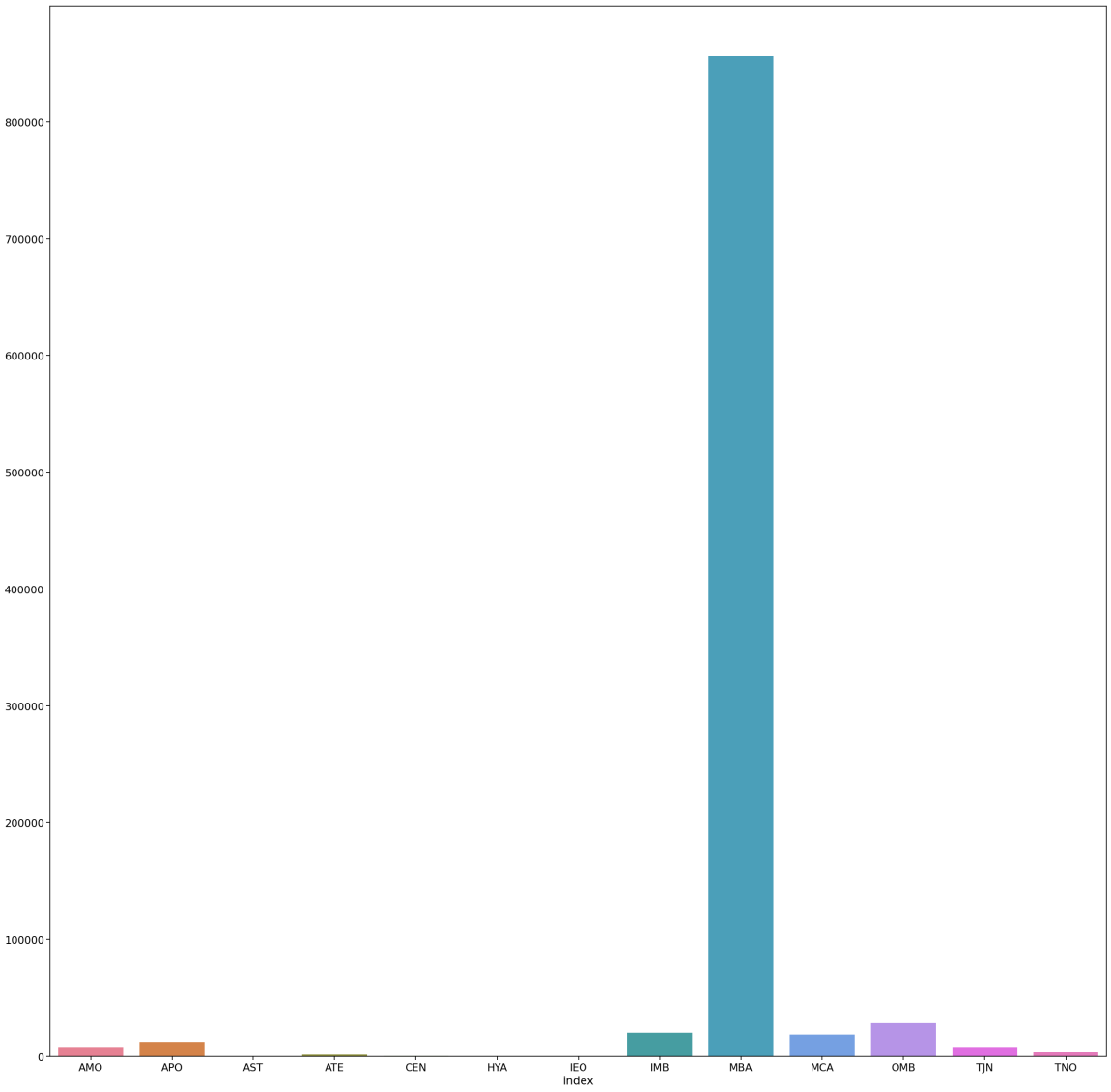


Figura 1.4: Frecuencia en cada clase de órbitas de asteroides.

a la clase MBA, es decir, gran cantidad de los asteroides tienen como característica que $2.0\ au < a < 3.2\ au$ y $q > 1.666\ au$. Esto tiene sentido ya que en el cinturón de asteroides es donde hay una mayor cantidad de ellos, este grupo de objetos se encuentra entre Marte y Júpiter (Figura 1.5).

- **Orbit_id.** Esta columna se utiliza para identificar y clasificar las órbitas de los asteroides. Cada elemento es un identificador único asociado a una órbita específica de un asteroide en particular. Esta columna es crucial para organizar y distinguir entre las diferentes trayectorias orbitales de los asteroides que están siendo registradas en la base de datos. La Tabla 1.1 contiene algunos elementos de esta columna. En particular, cada fila de D

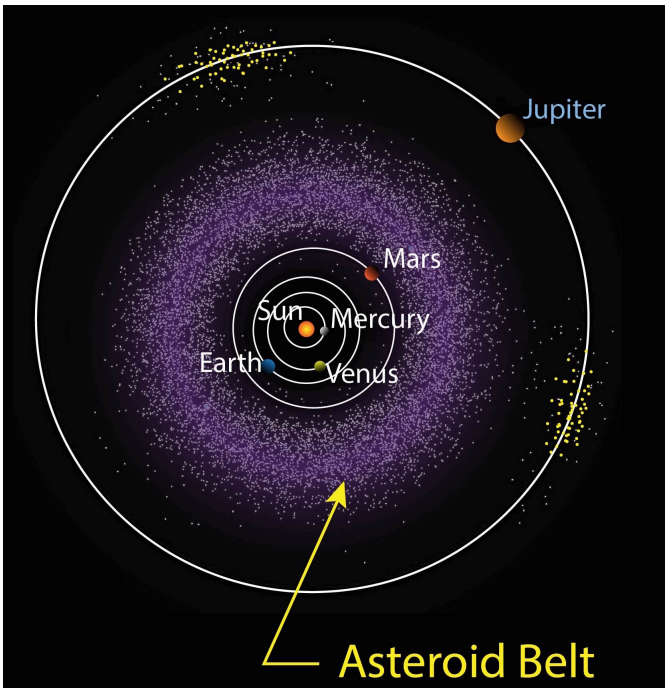


Figura 1.5: Cinturón de asteroides. Obtenida de [10].

contiene el identificador y la clase de órbita a la que pertenece el asteroide. Existen un

Id	órbita
a0000001	JPL 47
a0000002	JPL 37
a0000003	JPL 112
a0000004	JPL 35
bPLS6013	JPL 5
bT2S2060	JPL 3

Tabla 1.1: Clasificación de órbitas de algunos asteroides. En la primer columna se muestra el identificador único asignado a cada órbita de la segunda columna.

total de 4, 690 clases de órbitas, a cada una de ellas le corresponde al menos un asteroide.

Para la Figura 1.6, se tomaron las cincuenta órbitas con más frecuencias, se puede observar que hay mayor cantidad de asteroides en la órbita de clase 1, luego la clase JPL 1, JPL 2, y así sucesivamente. De aquí surge la siguiente pregunta: ¿Cuántos asteroides potencialmente peligrosos contiene cada órbita?, ya que no hay garantía de que las órbitas con más asteroides contengan la mayor cantidad de aquellos que son potencialmente peligrosos, es por ello que se requiere del análisis de las órbitas que contienen la mayor cantidad de objetos peligrosos. Obsérvese las tres barras más altas en la gráfica.

- **Orbit_id 1.**

  En la Tabla 1.5 se observa que de los 50,142 asteroides que se encuentran en la órbita 1, todos son clasificados como no potencialmente peligrosos.

- **Orbit_id JPL 1.** En la Tabla 1.3, de los 47,494 asteroides que se encuentran en

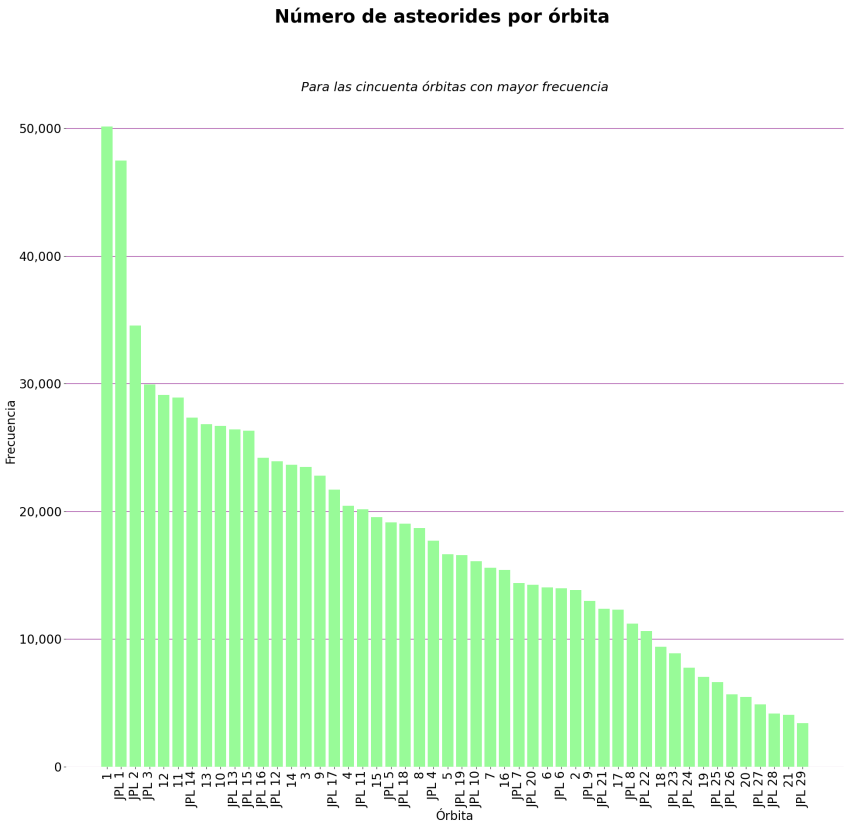


Figura 1.6: Cincuenta clases de órbitas con más frecuencias en la base de datos D.

1	
N	50,142
Y	0

Tabla 1.2: Número de asteroides pha en orbit_id 1.

la órbita llamada JPL 1, siete son clasificados como potencialmente peligrosos y se muestran en la Tabla 1.4.

En general, las órbitas con más frecuencias de pha con etiqueta “Y” se muestran en el gráfico de la Figura 1.7. En ella se observa que la clase de órbitas con mayor número de asteroides potencialmente peligrosos es la órbita llamada “18” y contiene exactamente 48 asteroides que son potencialmente peligrosos.

En la Figura 1.8 se pueden observar tres órbitas:

- Órbita terrestre en verde.
- Órbita de asteroide 543373 (2014 BJ56) en naranja, el cual pertenece a la clase N, es decir, no es potencialmente peligroso. Obsérvese la distancia relativamente grande respecto a la Tierra.
- Asteroide 2019 LH5. Clasificado como potencialmente peligroso. Note que su órbita intersecta a la terrestre en algunos puntos.

JPL 1	
N	47,487
Y	7

Tabla 1.3: Número de asteroides pha en orbit_id JPL 1.

JPL 1				
id	pha	class	a	q
bK19L05H	(2019 LH5)	APO	1.696463	0.472787
bK19L06G	(2019 LG6)	APO	2.277433	0.598619
bK19N02A	(2019 NA2)	APO	1.848841	0.662297
bK19Y06X	(2019 YX6)	APO	3.504298	0.380840
bK20B00X	(2020 BX)	APO	1.465895	0.866626
bK20C02B	(2020 CB2)	ATE	0.855635	0.615472
bK20F05A	(2020 FA5)	APO	1.943662	0.377813

Tabla 1.4: Asteroides pha en orbit_id JPL 1

En lo siguiente se muestran algunas características adicionales incluidas como variables en la base.

- **"Near-Earth Object" (NEO)**, que se traduce como Objeto Cercano a la Tierra. Es cualquier objeto celeste, como un asteroide o un cometa, que orbita dentro de la proximidad de la órbita terrestre. Estos objetos son de particular interés para los científicos y astrónomos debido a su proximidad y su potencial para impactar la Tierra.

  Para más información, es recomendable visitar la páginas web de la `https://cneos.jpl.nasa.gov/`, acerca de estos objetos cercanos a nuestro planeta.

- **Parámetro de magnitud absoluta (H).** En la astronomía de asteroides, el parámetro de magnitud absoluta (H) es una medida numérica que describe la magnitud aparente que tendría un asteroide si estuviera situado a una distancia estándar de 1 unidad astronómica del Sol y de la Tierra, y si su fase (ángulo de fase respecto a la órbita terrestre) fuera cero. Se utiliza como indicador de la luminosidad intrínseca del asteroide y es una herramienta importante para comparar y clasificar asteroides en función de su tamaño y brillo. En la Figura 1.9 se muestran las treintas magnitudes con más frecuencias.

- **Albedo.** De acuerdo con [8], albedo es el porcentaje de radiación que cualquier superficie refleja respecto a la radiación que incide sobre ella, véase la Figura 1.10 en la que se muestra esta forma de reflexión de la luz solar.

- **Albedo vs H.** Dado que la magnitud aparente disminuye a medida que aumenta la reflectividad (albedo) del objeto (ya que un objeto más reflectante parece más brillante), la relación entre el albedo y el parámetro de magnitud absoluta (H) es inversa: a mayor albedo, menor será el valor de H y, por lo tanto, más brillante será el objeto en magnitud aparente. Esta relación se muestra a través de la Figura 1.11.

JPL 2	
N	34,562
Y	6

Tabla 1.5: Número de asteroides pha en orbit_id JPL 2

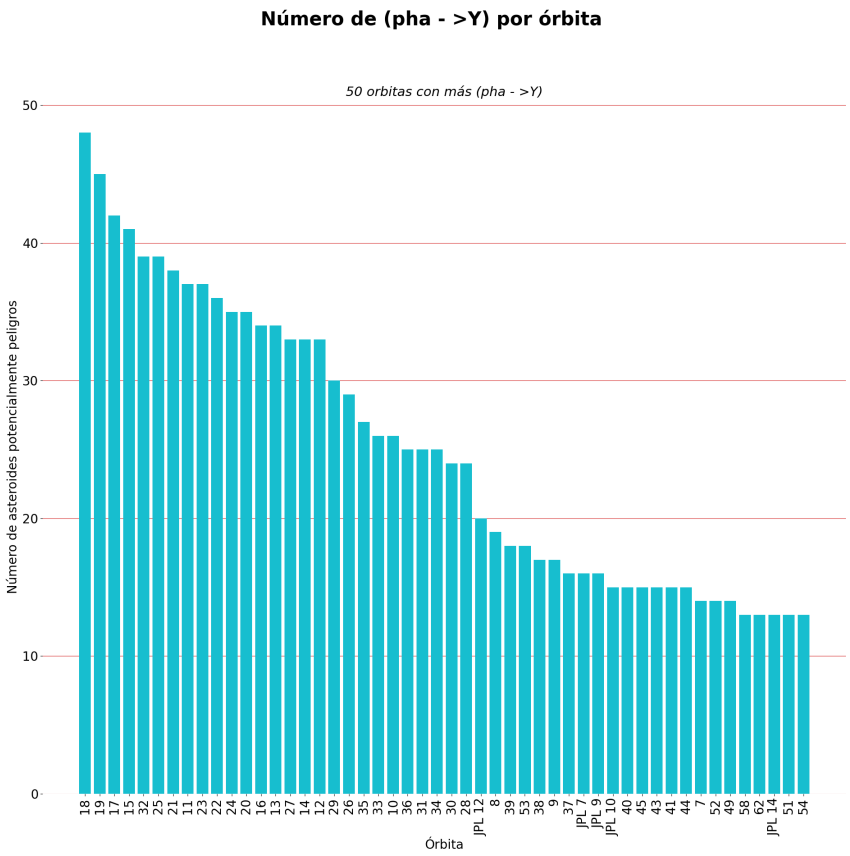


Figura 1.7: Cincuenta órbitas con más pha con etiqueta "Y" en D.

- **Epoch** El término "epoch" (época en español) se refiere a un punto específico en el tiempo que se utiliza como referencia para describir la posición o el estado de un objeto celeste en su órbita o movimiento aparente en el cielo. Esencialmente, la época proporciona un marco de referencia temporal para las observaciones astronómicas y los cálculos relacionados con el movimiento de objetos celestes.

  La elección de una época particular depende del objeto y del propósito de la observación o el cálculo. Por ejemplo, para los planetas y asteroides, la época puede ser una fecha específica en la que se determinó su posición orbital con precisión. Para objetos lejanos como estrellas o galaxias, la época puede ser una fecha de referencia estándar en el pasado o en el futuro, como el año 2000 o el año 3000.

  La información de la época es crucial para la predicción precisa de la posición futura de un objeto celeste en el cielo, así como para la comparación de observaciones realizadas en diferentes momentos en el tiempo. También se utiliza en la catalogación de objetos astronómicos y en la representación de sus órbitas en sistemas de coordenadas celestes.

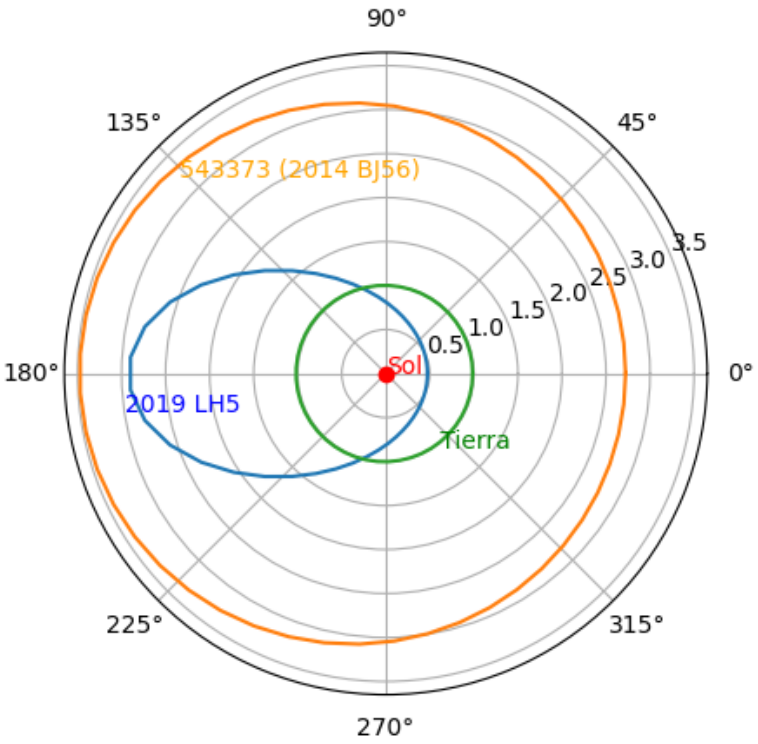


Figura 1.8: Órbitas de algunos asteroide y del planeta Tierra.

neo	
N	935,625
Y	22,895

- **H vs epoch.** La relación entre epoch y H se establece a través de la elección de la época específica que se utiliza como referencia para la magnitud absoluta (H) de un asteroide. En la Figura 1.12 se muestra que el valor H es constante para una gran mayoría de valores epoch, salvo algunos pocos casos, en la misma gráfica se muestra que para los diámetros altos se corresponde con valores de epoch bajos.

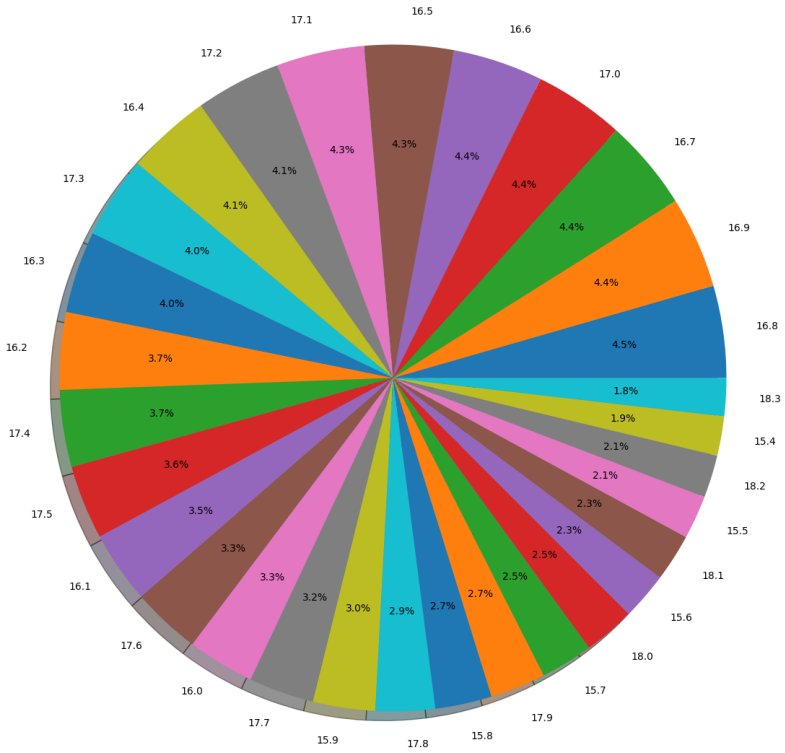


Figura 1.9: Parámetro de magnitud absoluta con mayor frecuencia de la base de datos D.

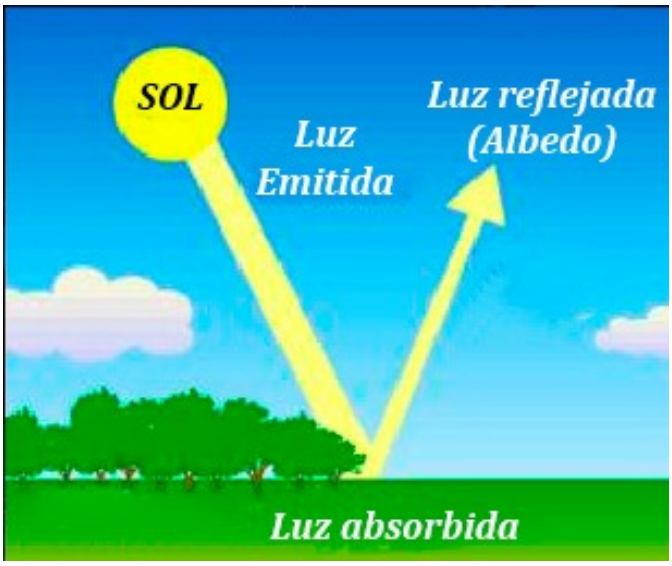


Figura 1.10: Albedo de la luz solar contra el suelo. Extraída de [8].

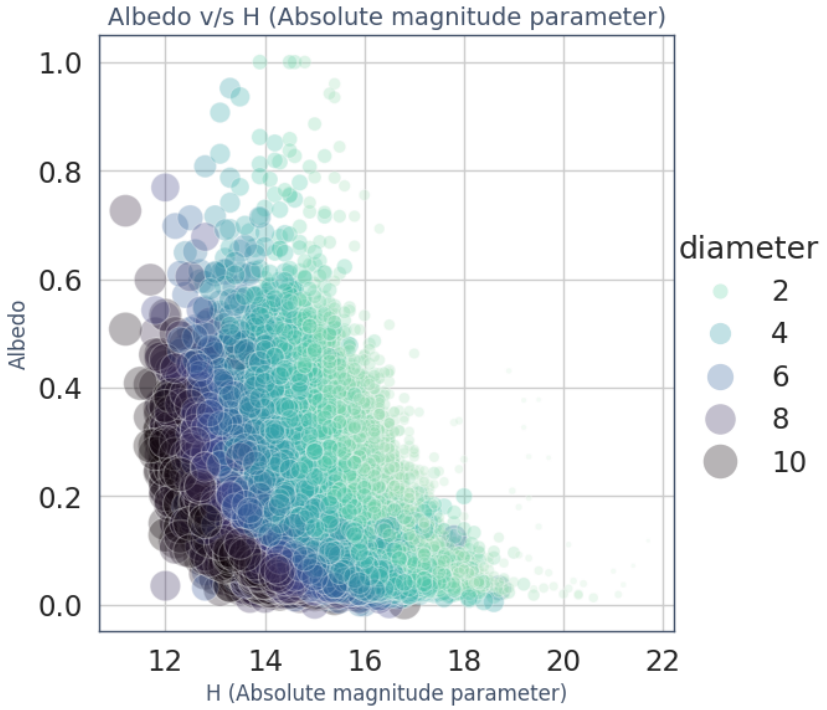


Figura 1.11: Relación entre H y albedo

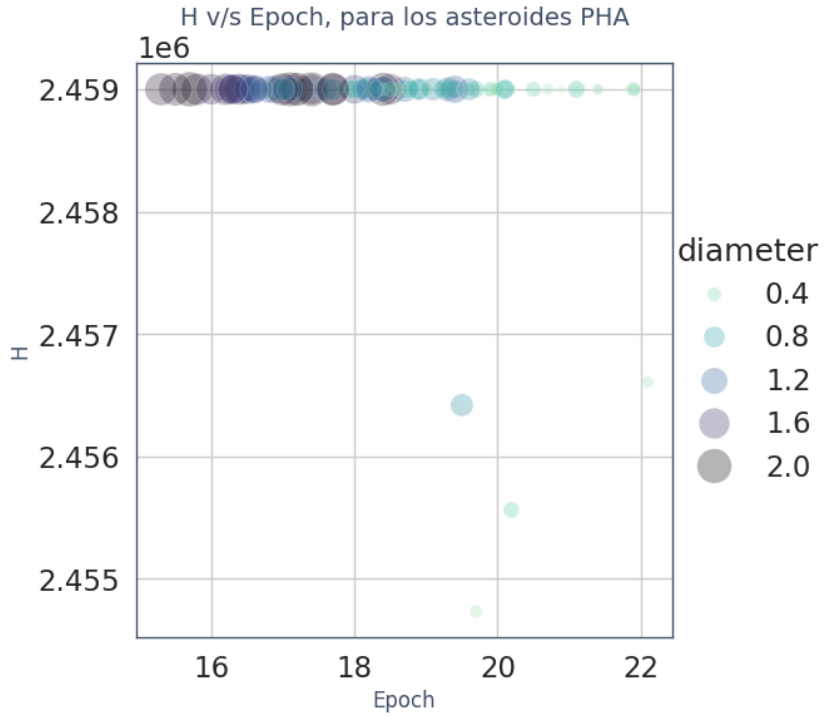


Figura 1.12: Relación entre H y epoch.

Capítulo 2

Imputación de datos

2.1. Datos perdidos y faltantes

Un dato perdido o faltante es aquel que se extravía en alguna etapa del procesamiento o creación de la base, o que no fue recopilado originalmente. En este trabajo de tesis serán utilizados ambos términos para referirnos a las entradas vacías debido a que se desconoce la clase a la que pertenece cada uno.

Como se mencionó en el capítulo anterior, la base D cuenta con datos faltantes de algunas características. En la Tabla 2.1 se exhiben los porcentajes de valores perdidos o faltantes.

Diversos autores recomiendan eliminar variables que tengan más del 25 % de valores faltantes, de hecho, en la literatura no existe un porcentaje aceptable de datos faltantes (véase por ejemplo [5]). En este sentido Schafer ([11]) menciona que un faltante del 5 % o menos no tiene efecto alguno en la inferencia estadística; Bennett ([2]) afirma que es probable que el análisis estadístico esté sesgado con una tasa de más del 10 % de faltantes. Más aún, la cantidad de datos faltantes no es el único criterio por el cual un investigador evalúa el problema, dentro de este contexto, Tabachnick y Fidell ([12]) postularon que los patrones generados por los datos faltantes tienen un mayor impacto en los resultados de la investigación que su misma proporción perdida. En [9, pág. 12], Medina y Galván mencionan de forma textual que *"no se recomienda imputar datos en situaciones en que la omisión en una o más variables alcance porcentajes superiores al 20 %"*. Por esta razón, se optó por eliminar las variables name, prefix, diameter, albedo y diameter_sigma debido a que más del 85 % son valores perdidos y, además, son características textuales que no son relevantes en la investigación. Por otro lado, se decidió imputar a las variables que presentan menos del 25 % de faltantes.

Al eliminar las columnas correspondientes a las variables, la Figura 1.3 del capitulo 1 se transforma en la mostrada en la Figura 2.1. Esta nueva base de datos será denotada por D_R.

2.2. Algoritmo EM

En esta sección se describirá el algoritmo de Expectation-Maximization (EM). Este algoritimo es una técnica estadística utilizada para encontrar estimaciones de máxima verosimilitud de parámetros en modelos probabilísticos, donde algunas de las variables son no observadas. Se utiliza comúnmente en problemas de aprendizaje, particularmente en situaciones donde hay

Características		Porcentaje
name	936460	97.69 %
prefix	958506	99.998122 %
neo	4	0.000417 %
pha	19921	2.078300 %
H	6263	0.653400 %
diameter	822315	85.789714 %
albedo	823421	85.905100 %
diameter_sigma	822443	85.803068 %
ma	1	0.000104 %
ad	4	0.000417 %
per	4	0.000417 %
per_y	1	0.000104 %
moid	19921	2.078300 %
moid_ld	127	0.013250 %
sigma_e	19922	2.078404 %
sigma_a	19922	2.078404 %
sigma_q	19922	2.078404 %
sigma_i	19922	2.078404 %
sigma_om	19922	2.078404 %
sigma_w	19922	2.078404 %
sigma_ma	19922	2.078404 %
sigma_ad	19926	2.078821 %
sigma_n	19922	2.078404 %
sigma_tp	19922	2.078404 %
sigma_per	19926	2.078821 %
rms	2	0.000209 %

Tabla 2.1: Porcentaje de datos perdidos por variable en D.

datos incompletos o valores faltantes. El algoritmo EM funciona en dos pasos principales:

- **Paso de Expectación (E):** En esta etapa, se calcula el valor esperado de las variables no observadas, dados los datos observados y las estimaciones actuales de los parámetros del modelo.
- **Paso de Maximización (M):** En esta etapa, se actualizan los parámetros para maximizar la verosimilitud de los datos observados, utilizando los valores esperados calculados en el paso de Expectación.

El proceso continúa de manera iterativa hasta que converge, es decir, hasta que los parámetros se estabilizan y no cambian significativamente entre las iteraciones. El algoritmo EM se utiliza ampliamente en diversos campos, incluyendo el aprendizaje automático, el procesamiento de señales, bioinformática, entre otros.

Al analizar la distribución de cada una de las columnas que serán imputadas, se observa que

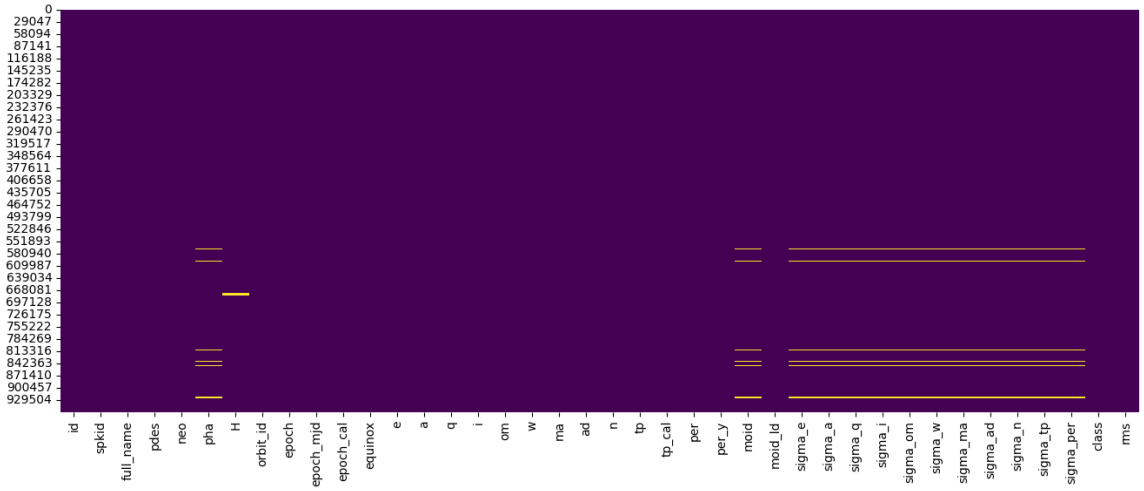


Figura 2.1: Mapa de calor de datos por imputar.

todas pertenecen a la familia exponencial, en la Figura 2.2, se muestra la distribución de **H, neo** y **pha**, respectivamente.

El algoritmo Expectation-Maximization (EM) fue propuesto de manera independiente por Arthur Dempster, Nan Laird y Donald Rubin (D. L. R.) en un artículo publicado en 1977 titulado "Maximum Likelihood from Incomplete Data via the EM Algorithm". Por lo tanto, se les atribuye la autoría y el nombre del algoritmo EM. Desde entonces, el algoritmo EM ha sido ampliamente utilizado y ha demostrado ser una herramienta poderosa en diversos campos de la estadística y el aprendizaje automático.

De acuerdo con D. L. R. el algoritmo puede ser simplificado cuando la distribución de los datos proviene de una familia exponencial, es decir, cuando la función de distribución es de la forma

$$f(X;\theta) = h(X)e^{\eta(\phi)T(X)-A(\phi)},$$

donde $h : \mathbb{R}^n \to \mathbb{R}$, $T : \mathbb{R}^n \to \mathbb{R}$, y tanto $\eta(\phi)$ como $A(\phi)$ son funciones reales definidas sobre el espacio de parámetros, todas ellas conocidas a priori. En este caso, ϕ es conocido como el parámetro de la familia. Es posible demostrar que algunos de los miembros de esta familia son la distribución de Poisson, Binomial, Normal, Gamma, Beta, entre otras.

Por otro lado, de acuerdo con [7], suponiendo que se tiene un conjunto de datos Y con una distribución de la familia exponencial y que $\phi_i^{(p)}$ es un estimador del parámetro ϕ después de p iteraciones del algoritmo EM; además, $Y = X \cup Z$, donde X representa a los datos observados y Z a aquellos que no son observados. Se definen formalmente los pasos **E** y **M** como sigue:

- **Paso-E:** Estimar el estadístico $T(Y)$ de los datos completos, calculando

$$T^{(p+1)} = E\left(T(Y)|\phi^{(p)}\right).$$

- **Paso-M:** Calcular ϕ^{p+1} como la solución del sistema

$$E\left(T(Y)|\phi\right) = T^{(P)}.$$

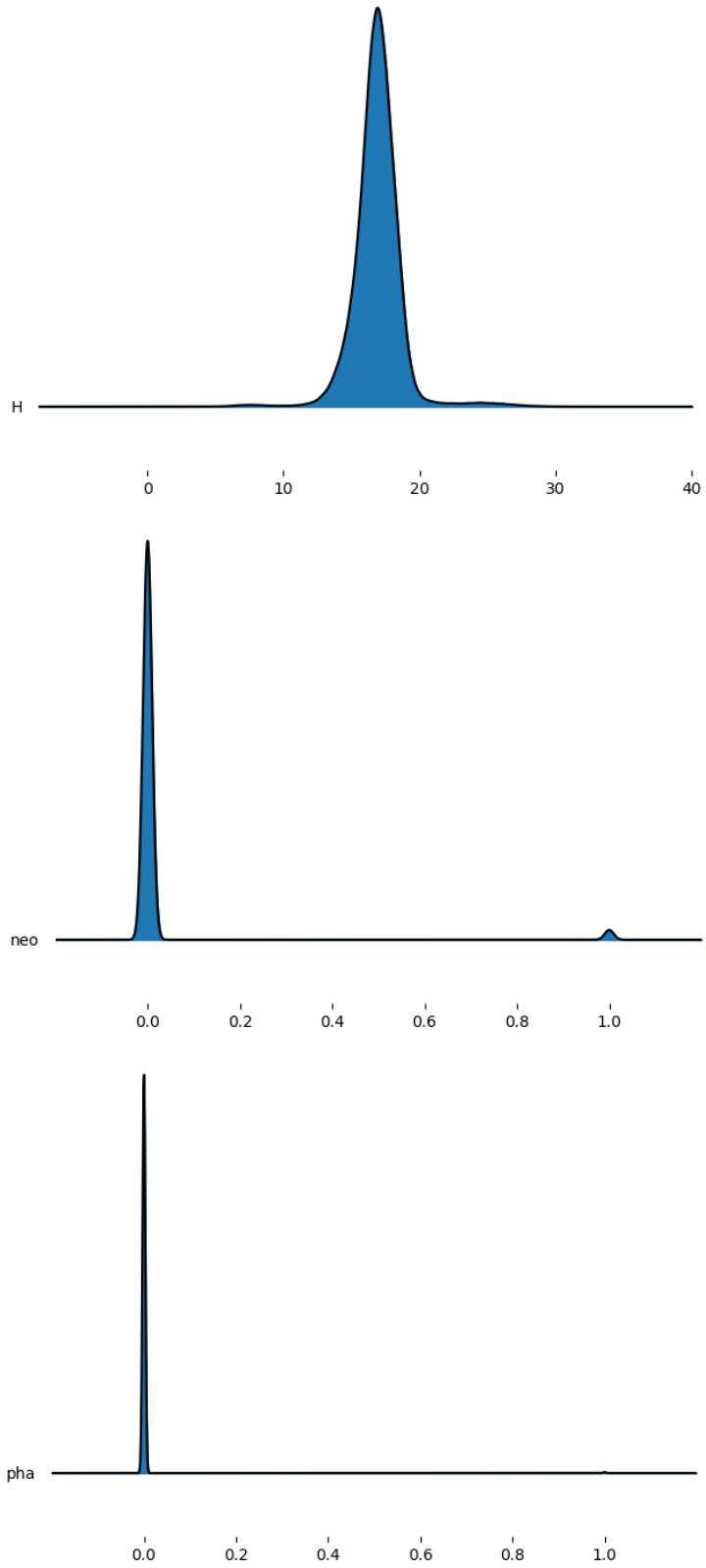


Figura 2.2: Distribución de las variables H, NEO y PHA, respectivamente.

Este sistema corresponde a la estimación por máxima verosimilitud de ϕ del conjunto de los datos completos Y, donde $T(Y)$ es reemplazado por $T^{(p)}$ en cada iteración del algoritmo.
Consideremos el siguiente ejemplo:
Supongamos que $[1, 2, 3, z_1, 4, z_2, 6] \sim N(\mu, \sigma^2)$, entonces $X = [1, 2, 3, 4, 6], Z = [z_1, z_2]$, sea $\phi_1^0 = 1$.

- **Paso E:**

$$T^{(1)} = E\left(T(Y)|\phi_1^{(0)}\right) = E\left([1, 2, 3, 4, 6]|1\right) = 2.83333.$$

- **Paso M:** Para el caso de la distribución normal, la media μ coincide con la máxima verosimilitud, así $\phi_1^{(1)} = 2.83333$.
- **Paso E:** $T^{(2)} = E\left(T(Y)|\phi_1^{(2)}\right) = E\left([1, 2, 3, 4, 6]|2.83333\right) = 3.13888$.
- **Paso M:** $\phi_1^{(1)} = 3.13888$.
- **Paso E:** $T^{(3)} = E\left([1, 2, 3, 4, 6]|3.13888\right) = 3.18981$.
- **Paso M:** $\phi_1^{(3)} = 3.18981$.
- **Paso E:** $T^{(4)} = E\left([1, 2, 3, 4, 6]|3.18981\right) = 3.1983$.
  $\vdots$
- $\phi_1^p \longrightarrow 3.2$, entonces $z_1 = 3.2$.

Ya que se tiene el valor estimado de z_1, se actualizan X y Y, esto es $X = [1, 2, 3, 3.2, 4, 6]$ y $Y = [z_2]$. Para estimar el valor de z_1, se aplica nuevamente el algoritmo EM, hasta que el estimador $\phi_2^{(p)}$ converge a un ϕ.

Debido a que las variables a imputar tienen distribuciones dentro de la familia exponencial es posible aplicar este proceso para la base de datos D_R con entradas vacías. El resultado se muestra en el diagrama de calor de la Figura 2.3, donde se observa que los datos faltantes han sido completados.

Esto conduce a que la base de datos esté lista para el siguiente paso (selección de características) que se analizará en el siguiente capítulo.

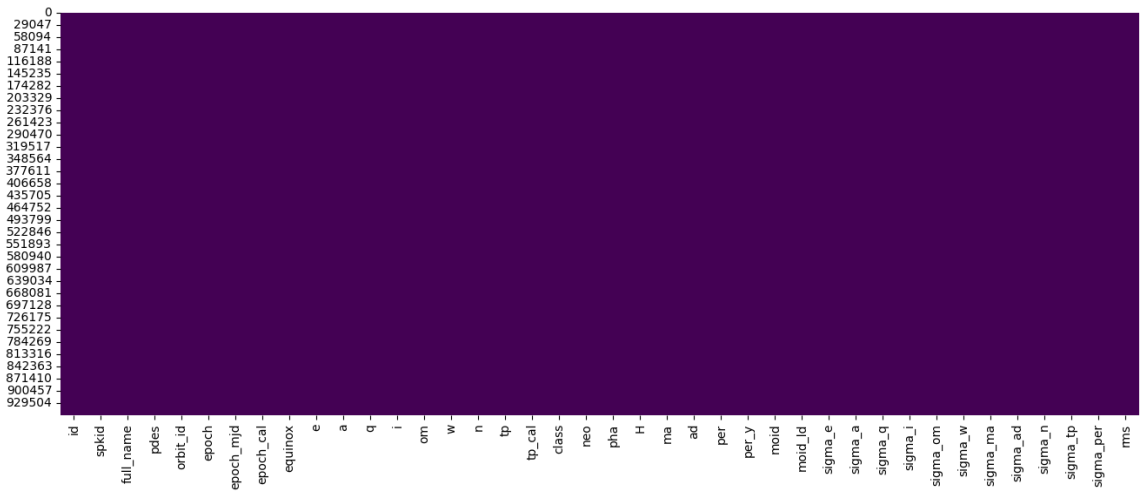


Figura 2.3: Mapa de calor de datos imputados.

Capítulo 3

Selección de características

La selección de características es el primer paso en el proceso de clasificación. Los datos reales pueden contener características de diversa relevancia para predecir etiquetas de clases. Por ejemplo, el sexo de una persona es menos importante que la edad para predecir los síntomas de enfermedades como la diabetes. Además de la ineficiencia computacional, las características irrelevantes a menudo afectan negativamente la precisión de los modelos de clasificación. Por lo tanto, el objetivo de los algoritmos de selección de característica es seleccionar las características más informativas con respecto a la etiqueta de clase. En este caso, se seleccionarán aquellas características que sean más relevantes para la clasificación de asteroides potencialmente peligrosos. Para ello es necesario que todas las características de la base de datos sean de tipo numérico. En la siguiente sección se muestran algunas técnicas que son capaces de convertir datos de tipo textual a numérico.

3.1. Codificación de variables

La codificación de variables es un proceso usual en la preparación de datos y el análisis estadístico. Se refiere a la transformación de variables categóricas (o cualitativas) en una forma numérica para que puedan ser incluidas en modelos estadísticos o algoritmos de aprendizaje automático.

Algunas técnicas comunes de codificación de variables son:

- **Codificación one-hot (dummy):** En esta técnica, cada categoría única en una variable categórica se convierte en una nueva variable binaria (0 o 1). Cada nueva variable representa la presencia o ausencia de esa categoría en la observación original. Es útil cuando las categorías no tienen un orden inherente.

- **Codificación ordinal:** Se utiliza cuando las categorías tienen un orden natural. Asigna valores numéricos a cada categoría de acuerdo con su posición en el orden. Por ejemplo, si se tiene una variable con categorías “bajo”, “medio” y “alto”, se asignan los valores 1, 2 y 3, respectivamente.

- **Codificación de frecuencias:** En esta técnica, las categorías se codifican según su frecuencia en los datos. Las categorías más frecuentes pueden recibir valores numéricos más bajos, mientras que las menos frecuentes pueden recibir valores más altos.

- **Codificación de target (o respuesta):** En esta técnica se asignan valores numéricos a las categorías basándose en la relación con la variable objetivo, es particularmente útil en problemas de clasificación. Por ejemplo, se asigna a cada categoría el promedio que toma la variable objetivo para esos registros en específico.

- **Codificación de efectos (o contrastes):** En esta técnica se establecen contrastes entre las categorías, esto es, se utilizan coeficientes para cada categoría que representan la diferencia entre esa categoría y una categoría de referencia.

La elección de la técnica de codificación depende del tipo de datos a analizar, el modelo y el objetivo del análisis. Es importante tener en cuenta el significado de las categorías y cómo se relacionan con la variable objetivo antes de seleccionar una técnica de codificación. Para el presente problema, se utiliza el algoritmo de minería **one-hot**, que funciona con datos numéricos en lugar de categóricos. La variable objetivo es *pha* en la base D_R, esta característica, como se mencionó anteriormente, etiqueta como Y a los asteroides potencialmente peligrosos y como N a los que no lo son. Por lo que es posible convertir estos atributos categóricos a formas binarias y entonces aplicar el algoritmo numérico. Formalmente, se introducen las siguientes definiciones para la codificación.

Definición 3.1. *Sea $\{Y, N\}$ el conjunto de clases de la variable pha. Se define la función de codificación como*

$$\phi(N) = -1;$$

$$\phi(Y) = 1.$$

Definición 3.2. *Sea neo el conjunto de objetos cercanos a la tierra o no (Y/N). Se define la función de codificación para este conjunto como*

$$\phi(N) = -1;$$

$$\phi(Y) = 1.$$

A este proceso se le conoce como portabilidad de los datos, en específico, binarización de atributos. Sin embargo, en la base de datos D_R, existen otras características que son de tipo categóricas que están compuestas de un número grande de clases, por lo que si se someten a la codificación one-hot se genera un número extremadamente grande de variables (una por cada clase). Para ilustrarlo, si se toma, por ejemplo, la columna *full_name*, cada entrada de esta columna es única, entonces al intentar hacer una codificación one-hot, solo para esta columna se tendría una matriz de tamaño $958,524 \times 958,524$. Es por ello que para estas columnas se aplicará el método de *codificación ordinal.*

Una escala de medición ordinal se logra cuando las observaciones pueden colocarse en un orden relativo con respecto a la característica que se evalúa, es decir, las categorías de datos están clasificadas u ordenadas de acuerdo con la característica especial que poseen ([4]). Por ejemplo, en la base de datos, se tienen algunas características de tipo categórico, como es el caso de: id, $full\_name$ entre otras. Se observa cómo después de la codificación, la medición ordinal toma por hecho que los elementos de cada columna ya están bien ordenados, así que los enumera

id	full_name
a0000001	1 Ceres
a0000002	2 Pallas
a0000003	3 Juno
a0000004	4 Vesta
a0000005	5 Astrea

Tabla 3.1: Antes de codificación

id	full_name
0.0	413389.0
1.0	413390.0
2.0	413391.0
3.0	413392.0
4.0	413393.0

Tabla 3.2: Después de codificación.

de forma ascendente como se muestra en la Tabla 3.2, en id, comienza desde 0, 1, 2, hasta n, en este caso n=958,523, ya que anteriormente se realizó una imputación así que todas las columnas tienen longitud n. Mientras que en $full\_name$ es similar, en este caso comienza en 413389.0, va a seguir de manera ascendente hasta cubrir el total de los elementos de la columna.

Ya que todas las características en la base de datos son de tipo numérico, esto nos permite pasar a la siguiente sección, a elegir las k mejores características, es decir, elegir las mejores características con mayor aportación a la investigación.

3.2. Modelo SelectKBest

Este estudio utilizó un modelo de paquete llamado SelectKBest es su abreviatura en inglés Select K best características o Selección de las k mejores características. Este modelo supone que se puede utilizar un algoritmo de clasificación para estimar qué tan bien el algoritmo maneja un subconjunto determinado de variables. Luego se utiliza un algoritmo de búsqueda de características en torno a este algoritmo para identificar conjuntos de características coincidentes. Los modelos de clasificación matemática son más precisos y tienen un conjunto óptimo de características, es decir, características que proporcionan la mayor información sobre las variables. En algunos casos, puede ser útil un algoritmo de clasificación específico para seleccionarlas. Por lo tanto, una de las entradas para la selección es un modelo de inducción, denotado por $\mathcal{A}$. Este tipo de algoritmo optimiza el proceso de selección de características, su estrategia básica es refinar de forma iterada a un conjunto de características Ω añadiéndole algunas nuevas de forma sucesiva. El algoritmo comienza inicializando el conjunto Ω como el conjunto $\emptyset$. La estrategia puede resumirse mediante los siguientes dos pasos que se ejecutan iterativamente:

1. Crear un conjunto aumentado de características Ω agregando una o más características al conjunto.

2. Usar el algoritmo de clasificación $\mathcal{A}$ para evaluar la precisión del conjunto de características Ω. Usar la precisión para aceptar o rechazar el aumento de Ω.

El aumento de Ω se puede realizar de muchas maneras distintas. Por ejemplo, se puede usar una estrategia donde el conjunto de características en la iteración anterior se aumenta con una característica adicional con el mayor poder discriminatorio con respecto a un criterio de filtro. Alternativamente, las características pueden seleccionarse para su adición a través de un muestreo aleatorio. La exactitud del algoritmo de clasificación $\mathcal{A}$ en el segundo paso se puede usar para determinar si se debe aceptar el conjunto de características recién aumentado o se debe volver al conjunto de características de la iteración anterior. Este enfoque se continúa hasta que no hay mejoría. Debido a que el algoritmo de clasificación $\mathcal{A}$ se usa en el segundo paso para la evaluación, el conjunto final de características identificadas será sensible a la elección del algoritmo $\mathcal{A}$.

3.2.1. Prueba ANOVA

El modelo de SelectKBest utiliza un algoritmo de filtrado $\mathcal{A}$ basado en los análisis de varianza (ANOVA). La técnica de análisis de varianza (ANOVA), constituye la herramienta básica para el estudio del efecto de uno o más factores (cada uno con dos o más niveles) sobre la media de una variable continua. Es por lo tanto, uno de los test estadísticos que se puede emplear para comparar las medias de dos o más grupos.

Para este proyecto, el objetivo es determinar si un asteroide es potencialmente peligroso o no lo es, es decir, es la variable *pha* de la base D_R, la cual es una variable discreta cuyas entradas toman valores -1 o 1, entonces, la hipótesis nula es que la media de esta variable es la misma que la media de cualquier otra de las variables, en contraposición a la hipótesis alternativa de que las medias difieren de forma significativa.

Se denomina factor a la variable que ejerce una influencia sobre la variable *pha*, a la que se denomina dependiente y, en lo siguiente, se realizará un análisis unifactorial que es definido a continuación.

Definición 3.3. *(Análisis de la varianza con un solo factor). Permite estudiar si existen diferencias significativas entre la media de una variable aleatoria continua en los diferentes niveles de otra variable cualitativa o factor, cuando los datos no están pareados. Es una extensión de los t-test independientes para más de dos grupos.*

Sin pérdida de generalidad, supongamos que la variable *pha* es la número d_R. Definamos μ al valor de su media y, para cada variable i-ésima, $i = 1, \ldots, d_R - 1$, denotemos como μ_i al valor de su media respectiva. Las hipótesis contrastadas en el ANOVA de un factor son:

- H_0 : No hay diferencias entre las medias, es decir,

$$\mu_i = \mu.$$

- H_1 : El par de medias son significativamente distintas la una de la otra.

Otra forma de plantear las hipótesis de un ANOVA es la siguiente: si se considera α_i el efecto introducido por el nivel i. La media de un determinado nivel (μ_i) se puede definir como:

$$\mu_i = \mu + \alpha_i.$$

- H_0 : El nivel no introduce un efecto sobre la media total: $\alpha_i = 0$
- H_1: El nivel introduce un efecto que desplaza su media: $\alpha_i \neq 0$.

En consecuencia, se toma la variable *pha* y se contrasta con cada una de las otras características, esto es, se realiza un ANOVA entre la variable *pha* y la variable i-ésima, si la hipótesis nula es aceptada, esta variable se rechaza y, en caso contrario, se agrega al conjunto Ω.
Para poder aceptar o rechazar la hipótesis nula, se determina el estadístico de Fisher, el cual está diseñado para atributos numéricos y mide el radio de la separación promedio entre clases al promedio de separación extra clase. Entre más grande es el estadístico mayor es el poder discriminatorio del atributo.

Sean $\mu_{i,j}$ y $\sigma_{i,j}$ la media y desviación estándar, respectivamente de los puntos que pertenecen a la clase $j \in \{m_{i_1}, m_{i_2}, \ldots, m_{i_n}\}$, donde m_{i_l} indica la $l-$ ésima clase de la $i-$ ésima variable para la característica i ordinal codificada y sea $p_{i,j}$ la fracción de puntos que pertenecen a la clase j. Recuérdese que μ_i es la media global de los datos sobre la característica i-ésima que está siendo evaluada.

Entonces, el estadístico de Fisher F_i para la característica i-ésima se define como:

$$F_i = \frac{\sum_{\{m_{i_1}, m_{i_2}, \ldots, m_{i_n}\}} p_{i,j}(\mu_{i,j} - \mu_i)^2}{\sum_{\{m_{i_1}, m_{i_2}, \ldots, m_{i_n}\}} p_{i,j}\sigma_{i,j}^2}.$$

El numerador del estadístico de Fisher cuantifica la diferencia media entre clases, mientras que el denominador cuantifica la diferencia media dentro de las clases. El atributo con el valor de puntuación de Fisher más alto se seleccionará para su uso en el modelo de la siguiente sección, esto debido a que la relación entre la variabilidad entre grupos y la variabilidad dentro del grupo sigue una distribución F cuando la hipótesis nula es verdadera. Por lo tanto, los valores F del estudio se colocan en la distribución F para determinar qué tan consistentes son los resultados con la hipótesis nula y para calcular la probabilidad de observar un estadístico F que sea al menos tan alto como el valor obtenido en la distribución F. Si la probabilidad es lo suficientemente baja, se concluye que los datos no respaldan la hipótesis nula, proporcionando evidencia de que la muestra es lo suficientemente fuerte como para rechazar la hipótesis nula para toda la población. Esta probabilidad también se llama valor p. Si los valores p de cada variable, que denotaremos como p_i, $i = 1, \ldots, d_R - 1$, son menores que 0.05, la variable será considerada y agregada al conjunto Ω, en caso contrario se desechará.

	Suma de cuadrados	Grados de libertad	F	Valor p
C (tratamientos)	3.006647e+07	1	3.854393e+07	0.0
Residual	1.495406e+06	1917046.0		

Tabla 3.3: Análisis ANOVA para las variables *pha* y *class*.

En el cuadro 3.3 se muestra el mismo análisis para las variables *pha* y *class*, obsérvese que

el valor p es muy pequeño, por lo que indica que es una variables considerada para pertenecer al conjunto Ω.

El método SelectKBest realiza el proceso ANOVA de todas las variables contra la variable objetivo *pha* y genera el conjunto Ω, luego toma las k mejores características, esto es, los k menores valores p, sin embargo, en el presente análisis resultan ser muy pequeños la mayoría de ellos. Una manera para seleccionarlos y observarlos es mediante un diagrama de barras en la que se grafican los valores $-\log_{10}(p_i)$, con $p_i \in \Omega$ (véase la Figura 3.1), de esta forma, las líneas más altas corresponden a las características significativas.

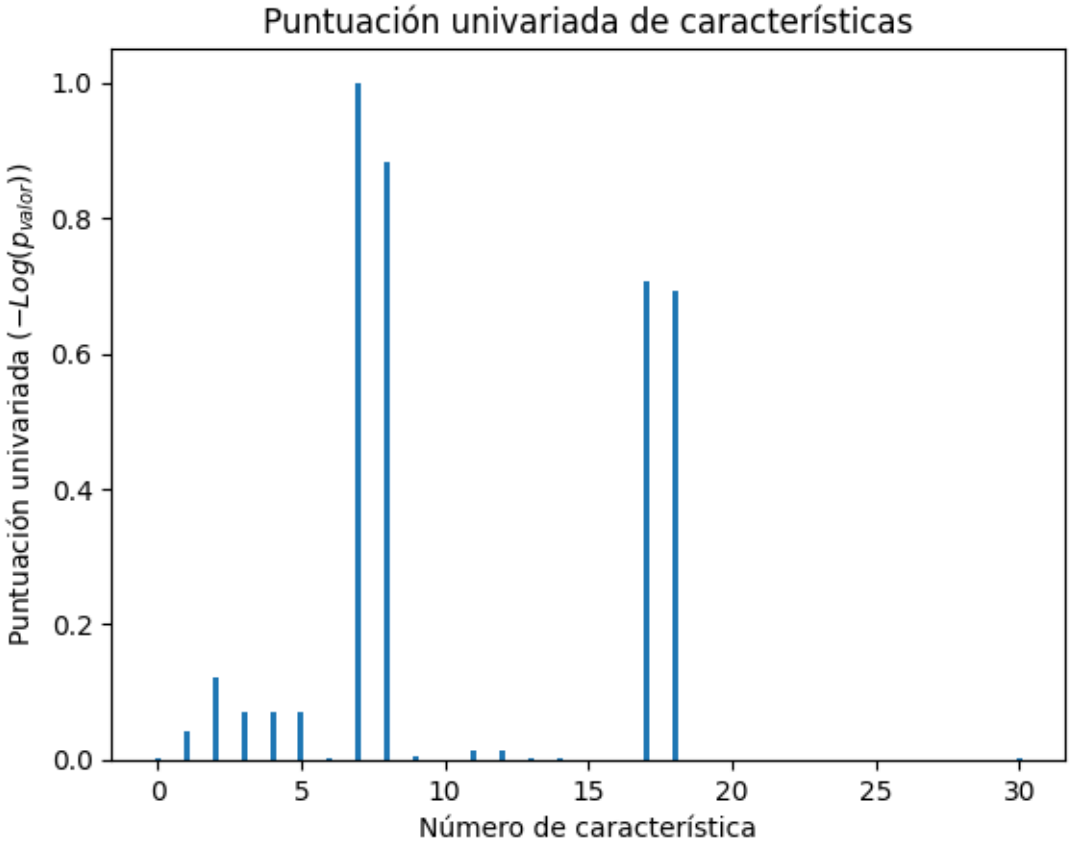


Figura 3.1: Gráfica de $-\log_{10}(p_i)$.

Por lo tanto, se consideran las nueve características con la barra más alta en la Figura 3.1, que corresponden a las variables:

- class: clase de órbita
- neo: objeto cercano a la tierra
- moid: distancia mínima de intersección de la órbita terrestre unidad au
- moid_id: distancia mínima de intersección de la órbita terrestre unidad lunar [1]
- H: Parámetro de magnitud absoluta

[1]Unidad Lunar o también conocido como LD es la distancia promedio desde el centro de la Tierra hasta el centro de la Luna que de acuerdo con la nasa equivale a 384 400 kilómetros.

- e: Excentricidad,
- q: Distancia del perihelio en *ua*
- n: Movimiento medio diario, es la velocidad angular de un astro en una órbita elíptica, medido en grados o radianes por día.
- i: inclinación.

Y este conjunto de variables es el ideal para realizar el entrenamiento del modelo, que será descrito en el capítulo siguiente.

Capítulo 4

Modelo de máquina de soporte vectorial

En este capítulo, se muestra el modelo que realiza la clasificación de nuevos registros $\hat{Y} \notin D_R$. Para ello, se divide la base D_R en una base compuesta por dos tercera parte de los registros en D_R, la cual se denotará como $D_{R,Train}$ y al número de filas por n^*; y el resto formarán un conjunto de registros para validación, denotada como $D_{R,Test}$. Por lo tanto, la base $D_{R,Train}$ será utilizada para entrenar el modelo basado en la teoría de máquinas de soporte vectorial.

Las máquinas de soporte vectorial (SVM) usualmente se definen para problemas de clasificación binaria de datos numéricos, aunque se extiende al contexto multiclase, tal extensión no será abordada en esta tesis. Las variables categóricas se transforman en datos binarios mediante el proceso de binarización dado por la operación de la Definición 3.1, en particular, las etiquetas de clase son elementos del conjunto $\{-1, 1\}$. Al igual que con todos los modelos lineales, las SVM utilizan hiperplanos de separación como límite de decisión entre las dos clases. En el caso de las SVM, el problema de optimización para determinar estos hiperplanos se plantea con la noción de margen.

Intuitivamente, un hiperplano de margen máximo es aquel que separa las dos clases, y para el cual existe una región (o margen) a cada lado del límite y que no contiene puntos de entrenamiento en él. Para entender este concepto, primero se discutirá el caso donde los datos son linealmente separables. En el cual, se construye un hiperplano lineal que separe exactamente los puntos pertenecientes a las dos clases. Por supuesto, este caso es inusual debido a que en la mayoría de las ocasiones los datos reales no se pueden separar por completo, y algunos puntos de datos mal etiquetados o valores atípicos no permitirán la separabilidad lineal. Por esta razón, la sección siguiente está dedicada a la formulación de la separación lineal.

4.1. SVM para datos linealmente separables

Como se mencionó anteriormente, los datos linealmente separables son aquellos que pueden ser divididos perfectamente por una línea (en dimensión dos), un plano (en dimensión tres), o un hiperplano (en dimensiones superiores, como en esta investigación), y los modelos SVM buscan maximizar el margen de separación entre las clases de tal manera que el hiperplano separador divida al espacio en dos regiones, donde los puntos de las diferentes clases queden en lados

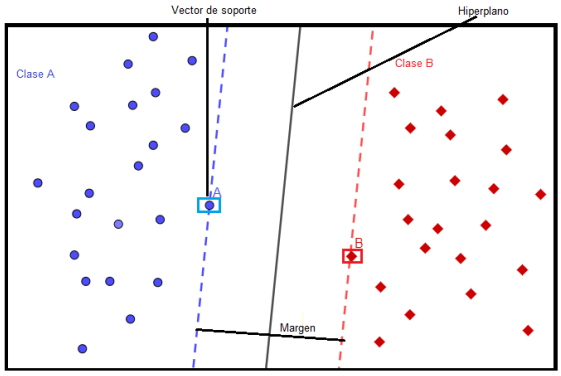


Figura 4.1: Datos linealmente separables

opuestos.

El margen es la distancia entre el hiperplano y los puntos más cercanos de cada clase (llamados vectores soporte), (obsérvese Figura(4.1)). Una SVM busca maximizar este margen, ya que un mayor margen implica una mayor confianza en la clasificación. Los vectores soporte son los puntos que están más cercanos al hiperplano separador y, por lo tanto, determinan su posición y orientación. Estos son cruciales debido a que el hiperplano óptimo se construye basándose en ellos.

¿Cómo se determina el hiperplano de margen máximo? Para hacerlo se plantea una fórmula de optimización no lineal, que maximice el margen expresándolo como una función de los coeficientes del hiperplano de separación. Al resolver el problema de optimización, se determinan los coeficientes óptimos. Se denotarán los n^* puntos en el conjunto de entrenamiento $D_{R,Train}$ por $(\overline{X}_1, y_1), \ldots, (\overline{X}_{n^*}, y_{n^*})$, donde $\overline{X}_i$ es un vector fila d-dimensional correspondiente al i-ésimo punto de datos, y $y_i \in \{-1, +1\}$ es la variable de la clase binaria del i-ésimo punto de datos. Entonces el hiperplano separador es de la siguiente forma:

$$\overline{W} \cdot \overline{X} + b = 0. \tag{4.1}$$

donde,

- $\overline{W} = (w_1, \ldots, w_{d_R})$ es el vector fila d_R-dimensional que representa la dirección normal al hiperplano, este vector regula la orientación del hiperplano,

- b es un escalar, también conocido como sesgo, y regula la distancia del hiperplano desde el origen.

Los coeficientes correspondientes a $\overline{W}$ y b deben estimarse de los puntos de entrenamiento para maximizar el margen de separación entre las dos clases. Debido a que se supone que las clases son linealmente separables, también se puede suponer que existe tal hiperplano. Todos los puntos $\overline{X_i}$ con $y_i = +1$ (clasificados como potencialmente peligrosos) estarán en un lado del hiperplano que satisfaga $\overline{W} \cdot \overline{X_i} + b > 0$. De manera similar, todos los puntos con $y_i = -1$ (clasificados como no peligrosos) estarán al otro lado del hiperplano satisfaciendo $\overline{W} \cdot \overline{X_i} + b < 0$. De aquí, se tiene

el sistema de ecuaciones siguiente:

$$\overline{W} \cdot \overline{X_i} + b \geq 0, \quad \forall i : y_i = 1, \tag{4.2}$$

$$\overline{W} \cdot \overline{X_i} + b \leq 0, \quad \forall i : y_i = -1. \tag{4.3}$$

Por otro lado, se puede suponer que el hiperplano de $\overline{W} \cdot \overline{X} + b = 0$ está ubicado en el centro de los dos márgenes que definen hiperplanos. Por lo tanto, los dos hiperplanos simétricos que tocan los vectores soporte (o también conocidos como puntos de entrenamientos) se puede expresar introduciendo otro parámetro c que regula la distancia entre ellos.

$$\overline{W} \cdot \overline{X_i} + b = +c, \tag{4.4}$$

$$\overline{W} \cdot \overline{X_i} + b = -c. \tag{4.5}$$

Es posible suponer, sin perder generalidad, que las variables $\overline{W}$ y b están apropiadamente escaladas, es decir, se dividen las ecuaciones entre c, de esta manera el valor de c se puede fijar como 1. Por lo tanto, los dos hiperplanos se pueden expresar de la siguiente forma:

$$\overline{W} \cdot \overline{X_i} + b = +1, \tag{4.6}$$

$$\overline{W} \cdot \overline{X_i} + b = -1. \tag{4.7}$$

Estas restricciones se denominan restricciones de margen. Los dos hiperplanos segmentan el espacio en tres regiones, por lo que se asume que ningún punto de entrenamiento se encuentra en la región entre estos dos hiperplanos, y todos los puntos de entrenamiento se asignan a una de las dos regiones externas restantes. Este se puede expresar como restricciones puntuales en los puntos de datos de entrenamientos de la siguiente manera.

$$\overline{W} \cdot \overline{X_i} + b \geq +1, \quad \forall i : y_i = +1, \tag{4.8}$$

$$\overline{W} \cdot \overline{X_i} + b \leq -1, \quad \forall i : y_i = -1. \tag{4.9}$$

Esto es equivalente a

$$y_i(\overline{W} \cdot \overline{X_i} + b) \geq +1, \quad \forall i. \tag{4.10}$$

El objetivo es maximizar este margen. Sean X_- y X_+ vectores soporte para la clase -1 y $+1$, respectivamente. Se sabe que la distancia mínima de un punto $\overline{X_i}$ a un hiperplano $\overline{W} \cdot X + b$ se calcula como:

$$d = \frac{|\overline{W} \cdot \overline{X_i} + b|}{||\overline{W}||}$$

Por lo que la distancia del vector X_- al hiperplano de separación es $d_- = -\frac{X_- \cdot \overline{W} + b}{||\overline{W}||}$. Dado que dicho hiperplano se encuentra ubicado en el centro de los márgenes, la distancia d_+ del vector

X_+ es igual a d_-. Por lo tanto, la longitud del margen es

$$\begin{aligned} d &= 2d_- \\ &= 2\frac{X_+ \cdot \overline{W} + b}{||\overline{W}||} \\ &= 2\frac{1}{||\overline{W}||}(X_+ \cdot \overline{W} + b) \\ &= 2\frac{1}{||\overline{W}||}(1) \\ &= \frac{2}{||\overline{W}||}. \end{aligned}$$

Sin embargo, maximizar $2/\|\overline{W}\|$ es equivalente a minimizar $\|\overline{W}\|^2/2$, este proceso esta sujeto al conjunto de restricciones lineales dadas por las Ecuaciones 4.8 y 4.9 en los puntos de entrenamiento. Es decir,

$$\begin{aligned} \text{mín} \quad & \|\overline{W}\|^2/2 \\ s.a \quad & \\ & y_i(\overline{W} \cdot \overline{X_i} + b) \geq +1, \quad \forall i : y_i \in \{-1, 1\}. \end{aligned}$$

Cada vector soporte conduce a una restricción, lo que tiende a hacer que el problema de optimización tenga un número relativamente grande de ellas y, como consecuencia, presente alta complejidad computacional. Esta clase de problemas de programación no lineal con restricciones se resuelve usando un método conocido como optimización lagrangiana.

La idea general es asociar un conjunto d_R-dimensional no negativo de multiplicadores lagrangianos $\lambda = (\lambda_1, \cdots, \lambda_{n^*}) \geq 0$ por las diferentes restricciones. El multiplicador λ_i corresponde a la restricción del i-ésimo vector soporte. Luego, la función objetivo se ajusta mediante la incorporación de una penalización lagrangiana para cada una de las restricciones:

$$L_P = \frac{\|\overline{W}\|^2}{2} - \sum_{i=1}^{n^*} \lambda_i[y_i(\overline{W} \cdot \overline{X_i} + b) - 1]. \tag{4.11}$$

Para valores fijos no negativos de λ_i, cualquier punto que se encuentre dentro del margen (conocida como violación de la restricción del margen) incrementa L_P. En consecuencia, el término de penalización λ fuerza que los valores optimizados de $\overline{W}$ y b minimicen el efecto de las violaciones de las restricciones, reduciendo L_P con respecto a $\overline{W}$ y b.

Los valores de $\overline{W}$ y b que satisfacen las restricciones de margen siempre darán como resultado una penalización no positiva. Por lo tanto, para cualquier valor fijo no negativo de λ, el valor mínimo de L_P siempre será a lo más $\frac{\|\overline{W^*}\|^2}{2}$, debido al impacto del término de penalización no positivo para cualquier punto factible (W^*, b^*). Por lo tanto, si L_P se minimiza con respecto a $\overline{W}$ y b para un $\overline{\lambda}$ en particular, y luego se maximiza con respecto a los multiplicadores lagrangianos no negativos $\overline{\lambda}$, la solución dual resultante L_D^* será un límite inferior de la función objetivo $O^* = \frac{\|\overline{W^*}\|^2}{2}$. Matemáticamente, esta condición de dualidad débil se puede expresar de la siguiente

forma:

$$O^* \geq L_D^* = max_{\overline{\lambda} \geq 0} min_{\overline{W},b} L_P. \tag{4.12}$$

Las formulaciones de optimización como SVM son especiales porque la función objetivo es convexa y las restricciones son lineales. Tales formulaciones satisfacen una propiedad conocida como dualidad fuerte. De acuerdo con esta propiedad, la relación minimax de la Ecuación (4.12) tiene como consecuencia que exista el punto óptimo y con él, se tenga una solución del problema original (es decir, $O^* = L_D^*$) en la que el término de penalización de Lagrange tiene una contribución cero. Tal solución se conoce como el punto de silla de la formulación de Lagrange. Tenga en cuenta que la penalización Lagrangiana cero se logra mediante una solución factible sólo cuando cada punto de entrenamiento $\overline{X}_i$ satisface $\lambda_i[y_i(\overline{W} \cdot \overline{X}_i + b) - 1] = 0$. Estas condiciones son equivalentes a las condiciones de optimización de Karush-Kuhn-Tucker [1] e implican que los puntos $\overline{X}_i$ con $\lambda_i > 0$ son vectores soporte. La formulación Lagrangiana se resuelve siguiendo los siguiente pasos:

1. **Simplificación del Objetivo Lagrangiano.** La función L_P se transforma en un problema de maximización pura al eliminar las variables de minimización $\overline{W}$ y b usando condiciones de optimización basadas en gradientes. Al igualar a 0 el gradiente de L_P con respecto a $\overline{W}$, se obtiene lo siguiente:

$$\begin{aligned} \nabla L_P = \nabla \frac{\|\overline{W}\|^2}{2} - \nabla \sum_{i=1}^{n^*} \lambda_i [y_i(\overline{W} \cdot \overline{X_i} + b) - 1] = 0 \\ \implies \overline{W} - \sum_{i=1}^{n^*} \lambda_i y_i \overline{X_i} = 0. \end{aligned} \tag{4.13}$$

   Por lo tanto, ya se cuenta con una expresión para $\overline{W}$ en términos de los multiplicadores de Lagrange y los vectores de soporte, es decir,

$$\overline{W} = \sum_{i=1}^{n^*} \lambda_i y_i \overline{X_i}. \tag{4.14}$$

   Además,

$$\frac{\partial L_p}{\partial b} = 0 \implies \sum_{i=1}^{n^*} \lambda_i y_i = 0.$$

2. **Construcción del Problema Dual:** La condición de optimización $\sum_{i=1}^{n^*} \lambda_i y_i = 0$ se

[1] Sea el problema $min\ f(x)$ sujeto a $f_i(x) \leq 0,\ i = 1, \cdots, m$, donde todas las funciones son diferenciables. Se dice que el punto $\overline{x} \in \mathbb{R}^n$, junto con el vector de multiplicadores $\overline{u} \in \mathbb{R}^n$, verifica las condiciones de Kuhn-Tucker del problema si

$$\begin{aligned} \nabla f(\overline{x}) + \sum_{i=1}^{m} \overline{u}_i \nabla f_i(\overline{x}) = 0 \\ f_i(\overline{x}) \leq 0,\ i = 1, \cdots, m \\ \overline{u}_i \geq 0,\ i = 1, \cdots, m \\ \overline{u}_i f_i(\overline{x}) = 0,\ i = 1, \cdots, m \end{aligned}$$

puede utilizar para eliminar el término $-b\sum_{i=1}^{n^*}\lambda_i y_i = 0$ de la función L_P. La Ecuación (4.14) se puede entonces sustituir en L_P para crear un problema dual L_D en términos únicamente de las variables de maximización $\overline{\lambda}$. Específicamente, la función objetivo para el problema dual Lagrangiano es la siguiente:

$$L_D = \sum_{i=1}^{n^*}\lambda_i - \frac{1}{2}\sum_{i=1}^{n^*}\sum_{j=1}^{n^*}\lambda_i\lambda_j y_i y_j \overline{X_i}\cdot\overline{X_j}. \tag{4.15}$$

El problema dual maximiza L_D sujeto a las restricciones $\lambda_i \geq 0$ y $\sum_{i=1}^{n^*}\lambda_i y_i = 0$. Note que la función L_D está expresada únicamente en términos de λ_i, las etiquetas de las clases y_j y los productos punto de parejas $\overline{X_i}\cdot\overline{X_j}$ entre los vectores soporte. Por lo tanto, para hallar los multiplicadores de Lagrange se requiere conocer sólo los valores de las variables de clase y de los productos escalares de los registros de las características $\overline{X_i}$.

3. **Estimación del valor de b.** El valor de b puede obtenerse a partir de las restricciones de la formulación original del modelo SVM, en las cuales los multiplicadores de Lagrange λ_r son estrictamente positivos. Para estos puntos de entrenamiento, la restricción de margen $y_r(\overline{W}\cdot\overline{X_r}+b) = 1$ se satisface exactamente según las condiciones de Kuhn-Tucker. Así, el valor de b puede derivarse de cualquier punto de entrenamiento $(\overline{X_r}, y_r)$ de la siguiente manera:

$$y_r\left[\overline{W}\cdot\overline{X}_r + b\right] = 1 \quad \forall r : \lambda_r > 0; \tag{4.16}$$

o equivalentemente a

$$y_r\left[\left(\sum_{i=1}^{n^*}\lambda_i y_i\overline{X}_i\cdot\overline{X}_r\right) + b\right] = 1 \quad \forall r : \lambda_r > 0. \tag{4.17}$$

La segunda relación se obtiene al sustituir la expresión de $\overline{W}$ en términos de los multiplicadores de Lagrange según la Ecuación 4.14. Esta relación se formula únicamente en función de los multiplicadores de Lagrange, las etiquetas de clase y los productos escalares entre las instancias de entrenamiento. El valor de b se puede encontrar a partir de esta ecuación y, para minimizar el error numérico, se puede promediar b sobre todos los vectores soporte con $\lambda_r > 0$.

4. **Definición de la Etiqueta de Clase para una Instancia de Prueba**: Para una instancia de prueba $\overline{Z}$, su etiqueta de clase $F(\overline{Z})$ se determina utilizando el límite de decisión, el cual se obtiene al sustituir $\overline{W}$ en términos de los multiplicadores de Lagrange (Ecuación 4.14). Esto se expresa de la siguiente manera:

$$F(\overline{Z}) = \text{sign}\left\{\sum_{i=1}^{n^*}\lambda_i y_i(\overline{X_i}\cdot\overline{Z}) + b\right\}. \tag{4.18}$$

Aquí, λ_i son los multiplicadores de Lagrange, y_i son las etiquetas de clase de los puntos de entrenamiento, $\overline{X_i}$ son los vectores de características de los datos de entrenamiento, $\overline{Z}$ es

el vector de características de la instancia de prueba, y b es el sesgo calculado previamente. Este límite de decisión permite clasificar la instancia de prueba $\overline{Z}$ basándose en su posición relativa al hiperplano de separación determinado por la SVM.

4.2. SVM para datos no separables

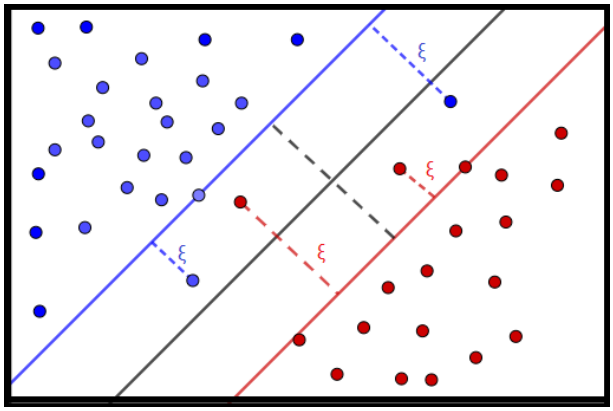


Figura 4.2: Datos no separables

En la sección anterior se examinó el caso en que los puntos de las dos clases son linealmente separables. Sin embargo, esta situación rara vez se encuentra en conjuntos de datos del mundo real. A pesar de esto, muchos conjuntos de datos reales son aproximadamente separables, lo que significa que la mayoría de los puntos pueden estar en los lados correctos de hiperplanos de separación bien elegidos. En estos casos, el concepto de margen se vuelve más flexible, permitiendo que algunos puntos de entrenamiento infrinjan las restricciones de margen a cambio de una penalización. Así, los dos hiperplanos de margen separan "la mayoría"de los puntos de entrenamiento, pero no todos como se muestra en la Figura (4.2).

El grado de violación de cada restricción de margen por parte del vector soporte X_i se denota mediante una variable de holgura $\xi_i \geq 0$. Así, el nuevo conjunto de restricciones suaves en los hiperplanos de separación se puede formular de la siguiente manera:

$$\begin{aligned} \overline{W} \cdot \overline{X}_i + b &\geq +1 - \xi_i, \quad \forall i : y_i = +1 \\ \overline{W} \cdot \overline{X}_i + b &\leq -1 - \xi_i, \quad \forall i : y_i = -1 \\ \xi_i &\geq 0, \quad \forall i. \end{aligned}$$

Estas variables de holgura ξ_i pueden interpretarse como las distancias de los vectores soporte al hiperplano de separación, cuando se encuentran en lado "incorrecto" del hiperplano. Los valores de ξ_i son ceros cuando los puntos están en el lado correcto del hiperplano de separación. No es deseable que demasiados vectores soportes tengan valores positivos de ξ_i por lo que tales violaciones son penalizadas mediante $C \cdot \xi_i^r$, donde C y r son parámetros definidos por el usuario que regulan el nivel de suavidad del modelo. Valores pequeños de C minimizan los errores en los puntos de entrenamiento, resultando en márgenes más estrechos. Configurar C lo suficientemente grande no permitiría errores en los puntos de entrenamiento para clases separables, lo cual equivale a establecer todas las variables de holgura en cero y utilizar la versión estricta del

problema. Una elección común para r es 1, conocido también como pérdida de bisagra. Por lo tanto, la función objetivo para SVM de margen suave, con pérdida de bisagra, se define de la siguiente manera:

$$O = \frac{||\overline{W}||^2}{2} + C\sum_{i=1}^{n^*} \xi_i. \tag{4.19}$$

Al igual que antes, este es un problema de optimización cuadrática convexa que puede resolverse mediante métodos lagrangianos. Se emplea un enfoque similar para formular la relajación lagrangiana del problema, incorporando términos de penalización y multiplicadores adicionales $\beta_i \geq 0$ para las restricciones de holgura $\xi_i \geq 0$:

$$L_p = \frac{||\overline{W}||^2}{2} + C\sum_{i=1}^{n^*} \xi_i - \sum_{i=1}^{n^*} \lambda_i \left[y_i \left(\overline{W} \cdot \overline{X}_i + b \right) - 1 + \xi_i \right] - \sum_{i=1}^{n^*} \beta_i \xi_i. \tag{4.20}$$

Se puede utilizar un enfoque similar al caso de margen recto de SVM para eliminar las variables de minimización $\overline{W}$, ξ_i y b de la formulación de optimización, creando una formulación dual de maximización. Esto se logra estableciendo los gradientes de L_P respecto a estas variables en 0. Al igualar a 0 el gradiente de L_P con respecto a $\overline{W}$, se obtiene lo siguiente:

$$\begin{aligned}
\nabla L_P &= \nabla \frac{\|\overline{W}\|^2}{2} + \nabla C\sum_{i=1}^{n^*} \xi_i - \nabla \sum_{i=1}^{n^*} \lambda_i [y_i(\overline{W} \cdot \overline{X_i} + b) - 1] - \nabla \sum_{i=1}^{n^*} \beta_i \xi_i = 0 \\
&\implies \overline{W} - \sum_{i=1}^{n^*} \lambda_i y_i \overline{X_i} = 0 \\
&\implies \overline{W} = \sum_{i=1}^{n^*} \lambda_i y_i \overline{X_i}.
\end{aligned}$$

Obsérvese que W es idéntico al caso de margen recto (Ecuación 4.14) y se cumple la misma restricción multiplicadora $\sum_{i=1}^{n} \lambda_i y_i = 0$. Los términos de holgura adicionales en L_P que involucran ξ_i no afectan los gradientes con respecto a $\overline{W}$ y b. La función objetivo L_D del dual lagrangiano en el caso de margen suave es idéntica a la del caso de margen recto (Ecuación 4.15), ya que los términos lineales que involucran ξ_i se evalúan como cero[2]. La única diferencia en el problema de optimización dual es que los multiplicadores de Lagrange no negativos satisfacen restricciones adicionales de la forma $C - \lambda_i = \beta_i \geq 0$, luego de establecer la derivada parcial de L_P con respecto a ξ_i como 0. Esto limita la influencia de cualquier punto $\overline{X_i}$ en el vector de peso $\overline{W} = \sum_{i=1}^{n^*} \lambda_i y_i \overline{X}_i$ a C debido a la suavidad del margen. El problema dual en SVM suave maximiza a L_D (Ecuación 4.15) sujeta a las restricciones $0 \leq \lambda_i \leq C$ y $\sum_{i=1}^{n^*} \lambda_i y_i \overline{X}_i = 0$. Las condiciones de optimización de Kuhn-Tucker para las restricciones de no negatividad de holgura son $\beta_i \xi_i = 0$. Como $\beta_i = C - \lambda_i$, obtenemos $(C - \lambda_i)\xi_i = 0$. Así, los puntos de entrenamiento X_i con $\lambda_i < C$ tienen $\xi_i = 0$ y pueden estar en el margen o en el lado correcto del margen. Los vectores soporte se definen como puntos que satisfacen exactamente las restricciones suaves de SVM, y algunos pueden tener una holgura distinta de cero. Los puntos con $\lambda_i > 0$ son siempre

[2]El término adicional en L_P que involucra a ξ_i es $(C - \beta_i - \lambda_i)\xi_i$. Este término se evalúa como 0 porque la derivada parcial de L_P con respecto a ξ_i es $(C - \beta_i - \lambda_i)$. Para que L_P sea óptimo, esta derivada parcial debe ser igual a 0.

vectores de soporte. Los vectores de soporte en el margen satisfacen $0 < \lambda_i < C$, útiles para resolver b. Para cualquier vector soporte X_r con holgura cero que satisfaga $0 < \lambda_r < C$, el valor de b se puede obtener como antes.

$$y_r \left[\left(\sum_{i=1}^{n^*} \lambda_i y_i \overline{X}_i \cdot \overline{X}_r \right) + b \right] = 1. \tag{4.21}$$

Esta expresión es similar a la del caso de SVM con margen recto, excepto que los puntos de entrenamiento relevantes se identifican con la condición $0 < \lambda_r < C$. La actualización de cada iteración en el método de ascenso de gradiente también es idéntica al caso separable (ver Sección 10.6.1.1 del libro [1], excepto que cualquier multiplicador λ_i que exceda C en cada iteración debe restablecerse a C. La clasificación de una instancia de prueba utiliza la Ecuacióm 4.18 en términos de los multiplicadores de Lagrange, ya que la relación entre el vector de peso y los multiplicadores es la misma en este caso. Por lo tanto, la formulación de SVM suave con pérdida de bisagra es muy similar a la formulación de SVM recta. Esta similitud es menor para otras funciones de penalización por holgura, como la pérdida cuadrática. La versión suave de SVM también permite una formulación primaria sin restricciones al eliminar las restricciones de margen y las variables de holgura simultáneamente, sustituyendo $\xi_i = \max\{0, 1 - y_i[W \cdot X_i + b]\}$ en la función objetivo primaria de la Ecuación 4.19. Esto da como resultado un problema de optimización sin restricciones en términos de W y b:

$$O = \frac{||\overline{W}||^2}{2} + C \sum_{i=1}^{n^*} \max\{0, 1 - y_i[\overline{W} \cdot \overline{X} + b]\}. \tag{4.22}$$

En muchos casos, los solucionadores lineales no son apropiados para problemas en los que el límite de decisión no es lineal. Para comprender este punto, considere la distribución de datos ilustrada en la Figura 4.3. Es evidente que ningún hiperplano de separación lineal puede delimitar las dos clases. Esto se debe a que las dos clases están separadas por el siguiente límite de decisión:

$$8(x_1 - 1)^2 + 50(x_2 - 2)^2 = 1. \tag{4.23}$$

Ahora, si uno ya tuviera alguna idea sobre la naturaleza del límite de decisión, podría transformar los datos de entrenamiento en el nuevo espacio de 4 dimensiones de la siguiente manera:

$$z_1 = x_1^2$$
$$z_2 = x_1$$
$$z_3 = x_2^2$$
$$z_4 = x_2.$$

El límite de decisión de la Ecuación (4.23) se puede expresar linealmente en términos de las variables $z_1, \ldots, z_4$, expandiendo la ecuación 4.23 en términos de x_1, x_1^2, x_2 y x_2^2

$$8x_1^2 - 16x_1 + 50x_2^2 - 200x_2 + 207 = 0$$
$$8z_1 - 16z_2 + 50z_3 - 200z_4 + 207 = 0.$$

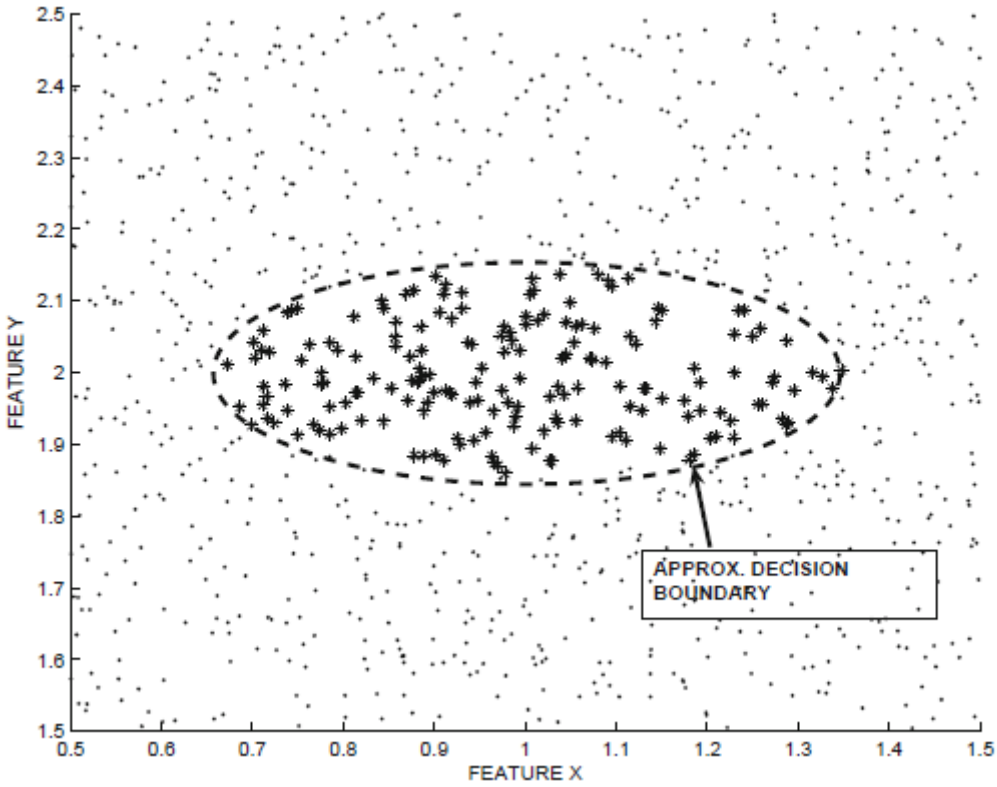


Figura 4.3: Superficie de decisión no lineal (Obtenido de [1])

Cada uno de los puntos de entrenamiento se transforma a un espacio de cuatro dimensiones, haciendo que las clases sean linealmente separables. La optimización SVM se resuelve como un modelo lineal en este nuevo espacio y se aplica también a instancias de prueba transformadas. Aunque esto incrementa la complejidad debido al mayor tamaño del vector de coeficientes del hiperplano W, permite aproximar límites de decisión polinomiales. Este método es útil cuando no se sabe si el límite de decisión es lineal o no lineal, ya que la flexibilidad del modelo puede adaptarse según los datos. Sin embargo, esta flexibilidad incrementa la complejidad computacional y el riesgo de sobreajuste si los datos de entrenamiento son insuficientes. Alternativamente, la "función de decisión"permite aprender límites de decisión no lineales sin transformar explícitamente el espacio.

4.3. Función de decisión.

El producto dentro del modelo de máquina de soporte vectorial puede ser sustituido por una función más general denominada kernel. El kernel se basa en la observación que la formulación SVM se puede resolver usando sólo los productos escalares (o similares) entre pares de puntos de datos, sin necesidad de conocer los valores individuales de las características. Así, la clave es definir el producto escalar por pares (o función de similitud) directamente en la representación transformada de d' -dimensiones $\Phi(X)$, utilizando una función kernel $K(X_i, X_j)$:

$$K(\overline{X}_i, \overline{X}_j) = \phi(\overline{X}_i) \cdot \phi(\overline{X}_j).$$

Para resolver eficazmente el SVM, no es necesario calcular explícitamente los valores de las características transformadas $\Phi(X)$, sino conocer el producto escalar (o similitud del kernel) $K(X_i, X_j)$. Así, el término $X_i \cdot X_j$ puede sustituirse por $K(X_i, X_j)$ en la Ecuación 4.15 y $X_i \cdot Z$ por $K(X_i, Z)$ en las ecuaciones, facilitando la clasificación SVM:

$$L_D = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \cdot \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j K(\overline{X}_i, \overline{X}_j), \tag{4.24}$$

$$F(\overline{Z}) = \text{sign} \left\{ \left(\sum_{i=1}^{n} \lambda_i y_i K(\overline{X}_i, \overline{Z}) \right) + b \right\}. \tag{4.25}$$

El sesgo b también se expresa en términos de productos escalares según la Ecuación 4.21.

Por lo tanto, todos los cálculos se realizan en el espacio original y no es necesario conocer la transformación real $\Phi(\cdot)$ siempre que se conozca la función de similitud del kernel $K(\cdot, \cdot)$. Utilizando kernel's seleccionados adecuadamente, se pueden aproximar límites de decisión no lineales arbitrarios. Existen diversas formas de modelar la similitud entre X_i y X_j, siendo algunas opciones comunes de funciones de kernel mostradas en la Tabla 4.1.

Función	
kernel lineal	$K(\overline{X}_i, \overline{X}_j) = \overline{X}_i \cdot \overline{X}_j$
Kernel de base radial gaussiana	$K(\overline{X}_i, \overline{X}_j) = e^{\|\|\overline{X}_i - \overline{X}_j\|\|^2 / 2\sigma^2}$
Kernel polinomial	$K(\overline{X}_i, \overline{X}_j) = (\overline{X}_i \cdot \overline{X}_j + c)^h$

Tabla 4.1: Tipos de kernel.

4.4. Métricas de validación.

Teniendo entrenado el modelo, la siguiente etapa es evaluarlo. La matriz de confusión es una herramienta útil para la evaluación del rendimiento de un modelo de clasificación. Está compuesta por cuatro componentes como se muestra en la Tabla 4.2.

		Clase predicha	
		-1	1
Clase Actual	-1	Verdadero Positivo (VP)	Falso Negativo (FN)
	1	Falso Positivo (FP)	Verdadero Negativo (VN)

Tabla 4.2: Matriz de confusión

En la matriz de confusión cada elemento corresponde a:

- Verdadero Positivo (VP): Aquí nuestro clasificador etiqueta un elemento positivo como positivo, lo que resulta en una victoria para el clasificador.
- Verdadero Negativo: Aquí el clasificador determina correctamente que un miembro de la clase negativa merece una etiqueta negativa.

- Falso Positivo (FP): El clasificador erróneamente llama a un elemento negativo como positivo, lo que resulta en un error de clasificación de "tipo I".
- Falso Negativo (FN): La clasificación declara erróneamente un elemento positivo como negativo, que da como resultado un error de clasificación de "tipo II".

A partir de la matriz de confusión, se pueden calcular varias métricas importantes para evaluar el desempeño de un modelo:

- **Accuracy (Exactitud)**: La proporción de predicciones correctas (tanto positivas como negativas) sobre el total de casos. Se calcula como:

$$\text{Accuracy} = \frac{VP + VN}{VP + FP + VN + FN}$$

- **Precision (Precisión)**: La proporción de verdaderos positivos sobre el total de predicciones positivas. Indica qué tan fiable es el modelo cuando predice un positivo. Se calcula como:

$$\text{Precision} = \frac{VP}{VP + FP}$$

- **Recall (Sensibilidad o Recuperación)**: La proporción de verdaderos positivos sobre el total de casos reales positivos. Mide la capacidad del modelo para identificar correctamente todos los positivos. Se calcula como:

$$\text{Recall} = \frac{VP}{VP + FN}$$

- **F1 Score**: Es la media armónica de la precision y recall, proporcionando una única métrica que balancea ambas. Es útil cuando se necesita un balance entre precision y recall. Se calcula como:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

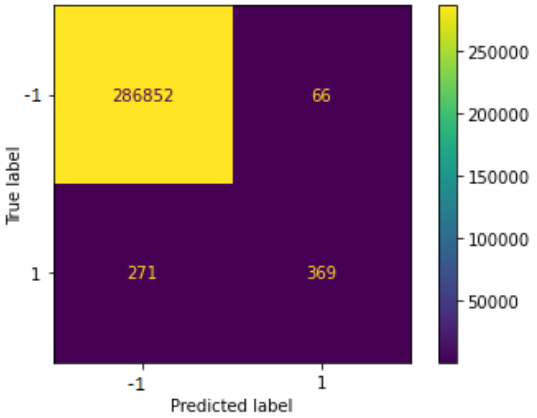
4.5. Resultados

En el proceso de entrenamiento y validación de los modelos de Máquina de Soporte Vectorial con diferentes kernels, se observaron comportamientos variados en cuanto a la convergencia y el desempeño. Para el modelo con kernel lineal, no se presentaron problemas durante el entrenamiento y la validación, logrando convergencia de manera eficiente. El modelo con kernel de base radial (RBF) también se entrenó sin inconvenientes, utilizando 150,000 iteraciones, alcanzando convergencia satisfactoria.

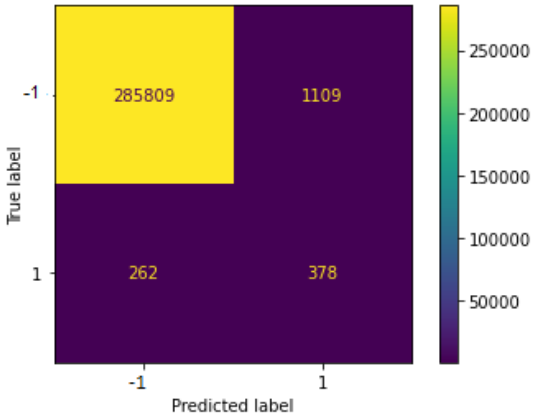
Sin embargo, el modelo con kernel polinomial presentó problemas de convergencia durante el entrenamiento. Inicialmente, se utilizó un polinomio de grado 3 con 150,000 iteraciones, pero no presentó cambios significativos en el ajuste al aumentar o disminuir el número de iteraciones. Para abordar este problema, se procedió a la estandarización de los datos, pero el problema persistió. Posteriormente, se incrementó el grado del polinomio a 4, lo que no solo no resolvió los problemas de convergencia, sino que también resultó en una disminución de las métricas de evaluación. Las métricas de cada modelo se presentan en la Tabla 4.3

Función	Accuracy	Precision	Recall	F1
Kernel lineal	0.9988281	0.84827586	0.5765625	0.686511628
Kernel rbf	0.9997566	0.96875	0.9203125	0.9439103
Kernel polinomial	0.9952323	0.2542031	0.590625	0.3554302

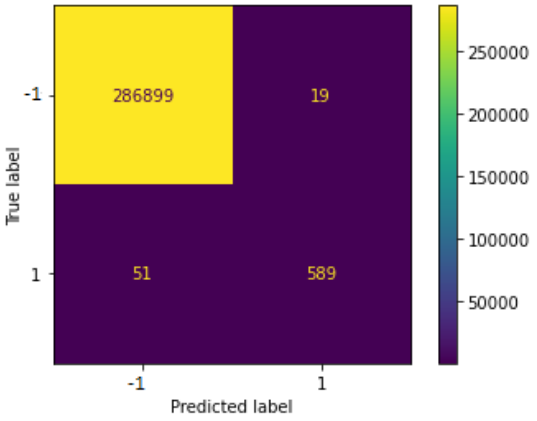
Tabla 4.3: Comparación de métricas entre los diferentes modelos SVM



(a) Matriz de confusión con kernel lineal.



(b) Matriz de confusión con kernel polinomial.



(c) Matriz de confusión con kernel rbf.

Figura 4.4: Métricas de evaluación.

Y la matriz de confusión de cada modelo se muestran en la Figura 4.4.

4.5.1. Análisis de la métricas de evaluación

- **Accuracy**: El modelo con Kernel RBF tiene la mayor exactitud ($Accuracy = 0.9997566$) (obsérvese la Tabla 4.3), seguido del modelo lineal y finalmente el polinomial. Esta métrica sugiere que el modelo RBF es ligeramente mejor en términos generales de clasificación correcta.

- **Precision**: Aquí, el modelo RBF nuevamente tiene la mayor precisión ($Precision = 0.96875$), indicando que, cuando el modelo predice que un asteroide es peligroso, es correcto el 96 % de las veces. Esto es crucial para minimizar los falsos positivos.

- **Recall:** El modelo RBF también tiene el mejor $recall$(0.9203125), lo que significa que detecta el 92 % de los asteroides realmente peligrosos, reduciendo así los falsos negativos.
- **F1 Score:** El modelo RBF alcanza un $F1Score$ de 0.9439103, indicando un balance óptimo entre precisión y recuperación.

El modelo SVM con kernel de base radial muestra un desempeño superior en todas las métricas evaluadas. Su alta $precision$ y $recall$ indican que es eficaz tanto en identificar asteroides potencialmente peligrosos como evitar falsas alarmas. Esto es fundamental para aplicaciones donde las consecuencias de clasificaciones incorrectas pueden ser significativas, como en la prevención y mitigación de impactos de asteroides.

El modelo lineal, aunque cercano en desempeño, muestra ligera desventaja en $recall$, lo que puede resultar en un mayor número de asteroides peligrosos no detectados. El modelo polinomial, si bien tiene un desempeño aceptable, no iguala al kernel RBF y el kernel lineal lo rebasa por $recall$.

Conclusión

El presente estudio sobre la clasificación de asteroides potencialmente peligrosos mediante el uso de una Máquina de Soporte Vectorial (SVM) ha permitido explorar y aplicar diversas técnicas de análisis y modelado de datos en cuatro capítulos fundamentales.

En el **Capítulo** 1, se realizó un análisis exploratorio y visualización de datos, donde se identificó una notable cantidad de datos faltantes en las características *name*, *prefix*, *diameter*, *albedo* y *diameter_sigma*, superando el 85 %. Esta observación subrayó la necesidad de una estrategia robusta para la imputación de datos, dado que la integridad de los datos es crucial para la eficacia de cualquier modelo predictivo.

En el **Capítulo** 2, se implementó el algoritmo de Expectación-Maximización (EM) para la imputación de datos. Se decidió eliminar las cinco características mencionadas en el Capítulo 1 debido a su elevado porcentaje de datos faltantes. Se observó que las columnas restantes con datos faltantes pertenecían a la familia exponencial, como fue propuesto por el autor mencionado. Esta característica simplificó los cálculos necesarios para la imputación. Posteriormente, los datos fueron preparados para la selección de variables, asegurando que solo se utilizaran las características más informativas. Además, se procedió a la codificación de las variables de tipo *object* mediante el método ordinal, facilitando así su inclusión en los modelos.

El **Capítulo** 3 se centró en la selección de características, utilizando el modelo *SelectKBest* para identificar las mejores variables predictivas. Tras la codificación de los datos, el modelo determinó que las nueve características más relevantes eran *class*, *neo*, *moid*, *moid_id*, *H*, *e*, *q*, *n* e *i*. Esta selección optimizó el conjunto de datos para el posterior entrenamiento del modelo SVM.

En el **Capítulo** 4, se desarrollaron y evaluaron los modelos de Máquina de Soporte Vectorial para datos linealmente separables y no separables. Se utilizaron los kernels lineal, polinomial y de base radial (RBF). Se analizaron las matrices de confusión y las métricas de validación para cada modelo, concluyendo que el kernel RBF presentó el mejor desempeño en comparación con los kernels lineal y polinomial. La evaluación detallada demostró que el modelo SVM con kernel RBF ofreció las métricas de precisión, recuperación y F1 Score más altas, posicionándose como la mejor opción para la clasificación de asteroides potencialmente peligrosos. Además, se concluye que los datos no son linealmente separables.

El código de programación utilizado en esta investigación se encuentra disponible en el enlace `https://acortar.link/KUuSQH` o en el siguiente código QR:

En resumen, este estudio no solo destacó la importancia de una adecuada imputación y selección de características, sino que también evidenció la superioridad del modelo SVM con kernel RBF para la tarea específica de clasificación, proporcionando una base sólida para futuras

investigaciones y aplicaciones en el campo de la astronomía y la seguridad espacial.

Bibliografía

[1] C. C. Aggarwal. *Data mining.* The McGraw-Hill Companies, Inc., 7 edition, 1997.

[2] D. A. Bennett. How can i deal with missing data in my study? *Australian and New Zealand journal of public health*, 25(5):464–469, 2001.

[3] Wikimedia commons. Lagrange points.jpg. `https://commons.wikimedia.org/wiki/File:Lagrange_points.jpg`, 2020.

[4] J Coronado Padilla. Escalas de medición. *Sistema Institucional de Investigación de Unitec (SIIU)*, 2007.

[5] Y. Dong and C. Y. J. Peng. Principled missing data methods for researchers. *SpringerPlus*, 2(222):1–17, 2013.

[6] M. Jiménez del Barco Ruiz-Herrera. ¿Qué son los puntos de Lagrange? `https://elseptimocielo.fundaciondescubre.es/descubre-el-universo/100-preguntas-100-respuestas/astronautica/que-son-los-puntos-de-lagrange/`, 2021.

[7] M. F. Lerdo Tejada Pavón. Estimación de datos faltantes con el algortimo em. *Tesis de licenciatura, Universidad Nacional Autónoma de México*, 2014.

[8] J. R. Lewy Soler. ¿Sabe usted que es Albedo? `https://astro.org.sv/publicaciones/sabe-usted-que-esalbedo/#:~:text=Los%20asteroides%20rojizos%20y%20de,las%20superficies%20clara%20la%20reflejan`, Julio 31, 2020.

[9] F. Medina and M. Galván. *Imputación de datos: Teoría y práctica.* Estudios estadísticos y prospectivos. Naciones Unidas, Santiago de Chile, julio de 2007.

[10] NASA Space Place. Asteroid or meteor: What's the difference? `https://spaceplace.nasa.gov/asteroid-or-meteor/en/`, January 25th, 2024.

[11] J. L. Schafer. Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1):3–15, 1999. PMID: 10347857.

[12] B. G. Tabachnick and L. S. Fidell. *Using multivariate statistics.* Allyn & Bacon, Needham Heights, MA, 6th edition, 2012.