Maestría en Explotación de Datos y Descubrimiento de Conocimiento

Materia: Análisis Inteligente de datos

Trabajo práctico: Asteroides Peligrosos

Alumno: Adrian Norberto Marino
Proyecto: aid-tp
RPubs: aid-tp

Introducción

En este trabajo práctico se aborda un problema de importancia como es la detección de objetos próximos a la tierra. Es un tópico de gran interés, ya que el estudio de los mismos permite predecir futuros desastres que podrían o no tomar vidas humanas o causar gran cantidad de heridos. Un ejemplo de esto es el incidente en Cheliábinsk (Rusia). El 15 de febrero del 2013 los habitantes de la ciudad de Cheliábinsk vieron en el cielo la estela de un asteroide de 20 metros de diámetro y más pesado que la torre Eiffel, el cual finalmente explotó en el aire. Todo el mundo se acercó a sus ventanas para ver este espectáculo natural. 90 segundos más tarde llegó la onda de choque (El sonido de la exploción) y las ventanas estallaron causando 1500 heridos y derrumbes en edificaciones.
La mayoría de los asteroides tienen órbitas estables entre Marte y Júpiter en el cinturón de asteroides principal. Algunos de estos objetos se encuentran muy cerca de la tierra, los cuales se denominan objetos próximos a la tierra. Es de gran interés estudiar estos objetos ya que tienen más posibilidades de impactar con la tierra. Estos asteroides son monitoreados por telescopios tanto en la tierra como en el espacio (Telescopio espacial Hubble). Estos telescopios toman fotos de distintas secciones del espacio periódicamente para detectar el movimiento de asteroides y otros cuerpos celestes. El problema de esta técnica, es que solo se puede detectar los objetos reflejados por el sol y aquellos que tienen gran tamaño. Dadas estas limitaciones es muy importante predecir con gran exactitud los asteroides peligrosos que sí podemos detectar. Por esta cuestión la NASA y otras organizaciones como la agencia espacial europea monitorean en forma constante estos objetos próximos a la tierra. En este trabajo práctico se abordará este problema e intentará predecir si un asteroide podría impactar con la tierra (Objeto peligroso).

Fuente de Datos

Para abordar este trabajo práctico se seleccionó el dataset NASA: Asteroids Classification. El mismo fue generado en el sitio cneos.jpl.nasa.gov el cual tiene una herramienta de consulta, tanto de asteroides como comentas. El dataset contiene 40 variables, en su mayoría cuantitativas (Continuas) y una pocas cualitativas (Categóricas y Nominales). Para más detalle ver nasa.csv.

Selección de variables

Para cumplir con las restricciones del trabajo práctico, las cuales se refieren a tomar 5 variables numéricas y una categórica, se realizaron los siguientes pasos:
Inicialmente se seleccionó la variable a clasificar (**Hazardous**). Esta es una variable dicotómica que nos dice si la observación (Asteroides) es peligrosa o no. Luego se excluyeron variables referidas a nombres e identificadores de cada objeto. También se excluyeron variables de serie de tiempo. Continuando, se filtraron las variables que tienen alta correacionas (>0.9) para evitar aquellas que fuesen colineales. Cabe aclarar que este dataset no tiene datos faltantes pero igualmente fueron omitidos en el pipeline de transformación de datos, ya que estas observaciones pueden causar

problemas con los algoritmos de clustering o clasificación. Luego se realizó la selección de variables probando 3 criterios distintos.

Métodos de selección

El primer método probado fue **feature importance** de Random Forest. Este método consiste en lo siguiente: Dado un conjunto de variables cualitativas y una categoría a predecir, se selecciona aquellas variables continuas que más ayudan en la predicción de los valores de la variable categórica a predecir.
Se aplicó el método sobre las variables predictoras(features) y se seleccionaron las variables más importantes según el índice **Mean Decrease Accuracy.** Este índice mide cuánta precisión pierde el modelo al quitar cada variable. Dicho esto, **feature importance** devuelve una lista de variables ordenada ascendentemente por pérdida de precisión.
Finalmente se tomó el top 5 de las variables con menor pérdida de presión resultando en la siguiente lista:

- Minimum.Orbit.Intersection
- Absolute.Magnitude
- Est.Dia.in.Miles.min
- Perihelion.Distance
- Inclination

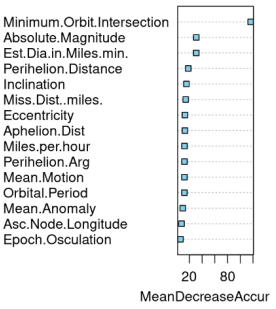

Figura 1. Orden de importancia de las variables.

Luego se realizó un **análisis de componentes principales(PCA)** para estudiar la correlación entre las variables originales e intentar determinar si había un contraste en ambas técnicas. A continuación podemos ver el biplot:

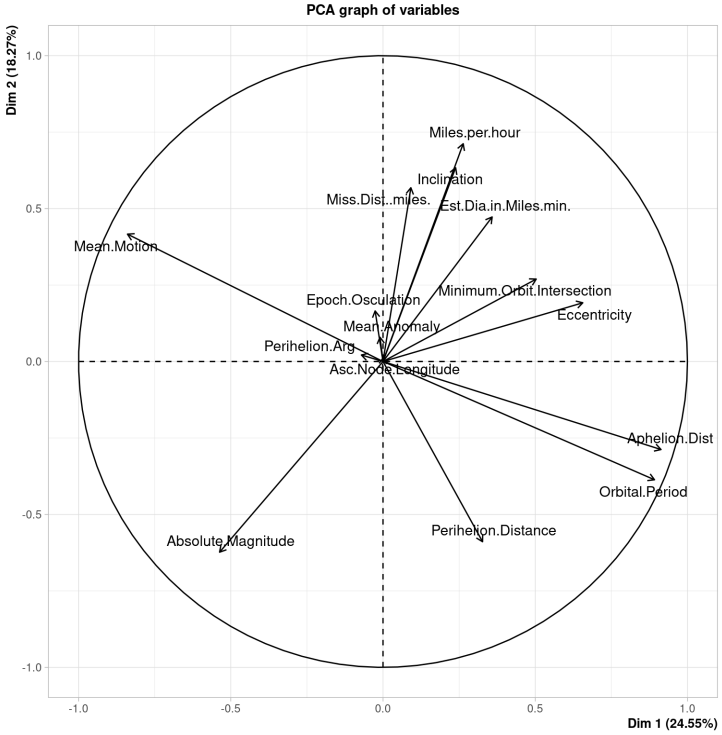

Figura 2. Biplot con variables originales.

A simple vista se aprecian 2 grandes grupos de variables correlacionadas tanto positiva como negativamente:

Grupo A	Grupo B
Absolute.Magnitude **Est.Dia.in.miles.min** **Minimum.Orbit.Intersection** **Inclination** Miles.pes.hour Miss.Dis.miles Eccentricity	**Perihelion.Distance** Mean.Motion Aphelion.Dist Orbital.Period

Nota: Solo te tuvieron en cuenta las variables con mayor varianza ya que podrían ser las que más información aportan al problema.

En negrita podemos ver el subgrupo de variables seleccionadas por **feature importance**. Todas las variables seleccionadas por **feature importance** tienen una gran componente de varianza en el biplot (figura 2). Finalmente se realizó un **cluster de variables** para tener otra alternativa más. Se seleccionó K=4 para el cual todavía el valor de Silhouette es considerablemente alto. Se encontraron los siguientes grupos:

Grupo 1	Grupo 2	Grupo 3	Grupo 4
Minimum.Orbit.Intersection **Inclination** **Est.Dia.in.Miles.min.** Miles.per.hour Miss.Dist..miles	**Absolute.Magnitude** Perihelion.Arg Mean.Motion	**Perihelion.Distance** Epoch.Osculation Asc.Node.Longitude Mean.Anomaly	Eccentricity Orbital.Period Aphelion.Dist

Nota: Nuevamente en negrita las variables seleccionadas por **feature importance**.

En la figura 1 se aprecia cierto grado de correlación entre las variables para el **Grupo 1**, ya que todas las variables tienen un ángulos menores a 40 grados (Aprox.) y son todas próximas. Lo mismo sucede en el **Grupo 3 y 4**, donde hay correlaciones positiva y negativas pero ninguna variable es ortogonal a la otra. En el **Grupo 2** tenemos variables que no tienen correlación entre ellas como **Absolute.Magnitude** y **Mean.Motion**.
En conclusión, no se encuentra un contraste claro entre las 3 técnicas para seleccionar las variables. Por esta cuestión se optó por el enfoque de **tomar las 3 técnicas como un hiper parámetro más de entrenamiento**. Es decir, se probará con la primera técnica **feature importance**, en caso de no tener buenos resultados se podría probar combinaciones de variables de los 4 grupos descubiertos por **clustering de variables** o tomar variables de gran varianza y baja correlación del biplot (figura 1).

Clasificación Supervisada

En este apartado vamos a buscar un modelo que nos permita predecir el valor de la variable **Hazardous,** sobre las observaciones que se encuentren en un conjunto de test extraído del dataset. La variable **Hazardous** define si una observación (asteroide) es peligrosa o no.

Selección de métrica de evaluación

Si contamos el número de observaciones por cada clase, encontramos que el dataset se encuentra muy desbalanceado:

- **No peligrosos**: 84% de las observaciones.
- **Peligrosos**: Solo el 15% de las observaciones.

Por otro lado, el costo es otro punto importante a tener en cuenta. Para este problema el mayor costo son las vidas humanas, y por esta cuestión un falso negativo tiene un gran costo. Un falso negativo, es un asteroide peligroso que el modelo clasifica como no peligroso.
En contraste, los falsos positivos tienen asociado un costo monetario, ya que se requiere tiempo de un experto para controlar los resultados del clasificar en caso de tener una alta tasa de falsos positivos.
Finalmente, se seleccionó **F beta score** como métricas de evaluación de los modelos con un valor **beta = 2**. De esta manera, priorizamos que el modelo capture el mayor número de verdaderos positivos, minimizando el número de falsos negativos.

Ejecución de Modelos

La clasificación se realizó con 5 modelos distintos: LDA, QDA, RDA, SVM, Regresión Logisticas y XGBoost.
Para aplicar LDA, QRD o RDA es necesario analizar los supuestos de normalidad y homocedasticidad multivariada de las 5 variables seleccionadas. Los resultados obtenidos fueron lo siguientes:

- **Test de shapiro-Wilk (Normalidad multivariada):** El p-valor < 0.05, por lo tanto hay evidencia para rechazar la hipótesis nula de normalidad y se rechaza normalidad multivariada.
- **Test de homocedasticidad multivariada (Box's M):** El p-valor < 0.05 por lo tanto se rechaza la hipótesis nula y podemos decir que las variables no son homocedásticas.

Finalmente, en este caso no es válido aplicar LDA, ya que no se cumplen ambos supuestos. Tampoco sería válido aplicar QDA/RDA, ya que no se cumple el supuesto de normalidad. Además hay una importante presencia de outliers, lo cual es otra contra para aplicar LDA. A pesar de esto, se decidió igualmente utilizar los modelos, ya que en algunos casos es posible tener resultados aceptables.

El umbral de selección de las clases es otro punto a tener en cuenta. Es el valor desde el cual se determina si la clase es **Peligroso** o **No-Peligroso**. Todos los modelos por defecto tienen un umbral de 0.5. En este caso se aplicaron 3 estrategias distintas, las cuales se tomaron como un hiper parámetro más:

- **Estrategia 1:** Seleccionar el umbral que minimiza los Falsos Positivos.
- **Estrategia 2:** Seleccionar el umbral que maximiza el AUR (Área bajo la curva ROC).
- **Estrategia 3**: Fijar el umbral en 0.5 para todos los modelos.

Resultados

Estrategia 1: Minimizar FN

Falsos Negativos	Falsos Positivos	Test F2Score %	Train F2Score %	Umbral	Modelo
0	63	90.498	88.485	0.01	QDA
0	66	90.090	88.576	0.01	RDA
0	279	68.259	66.149	0.01	SVM
0	364	62.241	60.976	0.01	LDA
1	4	98.673	99.438	0.01	XGBoost
15	12	87.940	86.379	0.01	Regresión logísticas

Estrategia 2: Maximizar AUR

Falsos Negativos	Falsos Positivos	Test F2Score %	Train F2Score %	Umbral	Modelo
1	4	98.673	99.438	0.01	XGBoost
1	14	97.064	96.610	0.38	QDA
1	14	97.064	96.692	0.39	RDA
3	33	92.857	92.231	0.28	SVM
3	68	87.970	88.073	0.26	LDA
15	11	88.087	86.455	0.04	Regresión logísticas

Estrategia 3: Umbral = 0.5

Falsos Negativos	Falsos Positivos	Test F2Score %	Train F2Score %	Umbral	Modelo
3	4	97.338	100.00	0.5	XGBoost
4	9	95.868	94.476	0.5	QDA
4	9	95.868	94.476	0.5	RDA
18	7	86.587	83.518	0.5	Regresión logísticas
28	4	79.861	80.769	0.5	SVM
36	21	71.795	72.321	0.5	LDA

Selección del mejor modelo

Vistos los resultados y comprendiendo los costos asociados a FN y FP, es comprensible buscar un tradeoff entre ambos valores para tener el mínimo número de FN y a su vez un número no muy alto FP.
A primera vista uno podría seleccionar el modelo QDA con la estrategia 1. Este modelo tiene cero FN y el menor FP dentro de los que tienen cero FN. El problema con este modelo es que el costo de FP es muy elevado. Pensemos en una detección diaria; deberíamos tener a un experto realizando un análisis manual de 63 objetos. Por esta cuestión, es más comprensible buscar el modelo que minimice ambos

costos dando más importancia a los FN. Entendiendo este criterio, el mejor modelo para la estrategia 1 sería XGBoost. Con el mismo criterio también seleccionamos XGBoost para la estrategia 2 y 3. Notemos que para XGBoost con las estrategias 1 y 2 se seleccionó el mismo umbral de corte.

Figura A: Se grafica un biplot con las clases FP, FN, TP, PN para el modelo seleccionado. Se puede apreciar que el modelo no predice bien (FN y FP) dentro del área donde se encuentran las observaciones positivas.

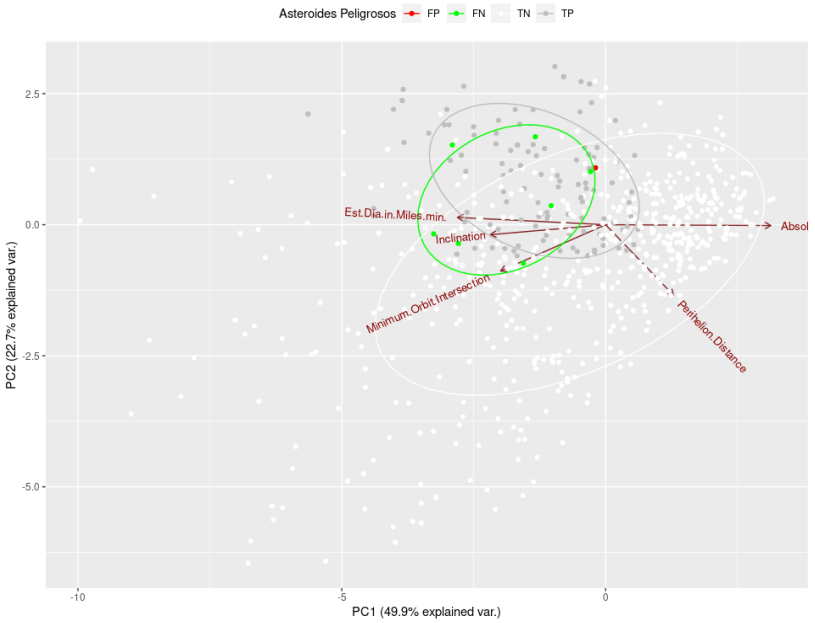

Figura B: Ide a Figura A pero para el segundo mejor modelo QDA con máxima AUR. (En las Figuras A y B puede que no coincidan exactamente las observaciones FN y FP con las tablas, ya que se generaron en otra corrida del algoritmo y difieren los numero pero el comportamiento en general es el mismo).

Finalmente, se seleccionó el modelo XGBoost con el umbra 0.01, en el cual tenemos el menor número de FN, un número aceptable de FP y el máximo AUR.

Clasificación No Supervisada

Los métodos no supervisados a diferencia de los supervisados, no requiere de un label para el entrenamiento, sino que descubren los labels en base a un conjunto de variables. Es decir, agrupa los individuos en grupos que se pueden intersecar o no, y a cada grupo le podemos asignar uno o más labels.
En este apartado, la idea es descubrir grupos que tengan el menor solapamiento posible, comprender qué características tienen los individuos que se encuentran en cada grupo y si es posible, nombrar los grupos. Para realizar esta tarea, vamos a utilizar el algoritmo k-means.

K-Means

Antes de comenzar con k-means, es necesario determinar cuál es el número óptimo (K) de clusters a descubrir, dado que este es un hiper parámetro del modelo. Para determinar el parámetro K se seleccionó la métrica Silhouette. El método Silhouette calcula la silueta promedio de observaciones para diferentes valores de K. Cuanto más alto sea el valor de Silhouette, más cohesión habrá entre los grupos y por lo tanto, más separados unos de otros. A continuación se expone el cálculo de Silhouette para dos normalizaciones diferentes: z-score y min-max:

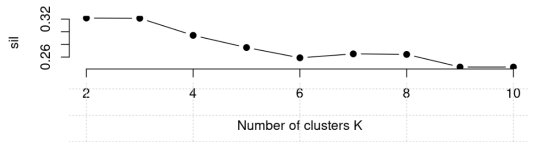

Figura 3. Silhouette vs. K con variables previamente normalizadas con z-score scale.

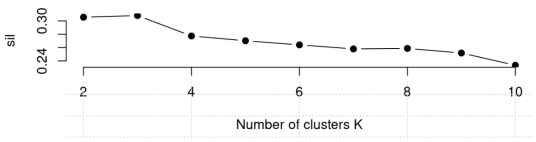

Figura 4. Silhouette vs. K con variables previamente normalizadas con min-max scale.

De los resultados de las figuras 3 y 4 se seleccionó la normalización min-max con K=3 por tener el mayor valor de Silhouette. Luego de aplicar el algoritmo k-means, se grafico un biplot utilizando PCA robusto (Usando MVE) para visualizar los grupos. MVE permite minimizar los outliers en el biplot, ya que algunas observaciones no permiten ver los grupos correctamente.

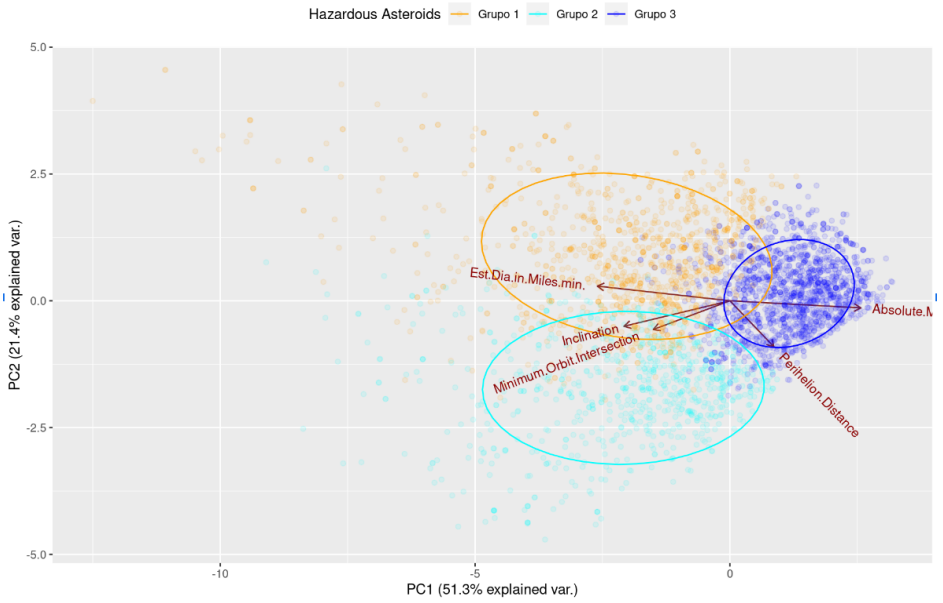

Figura 5. Visualización de grupos descubiertos por k-means en un biplot.

En el biplot se puede apreciar que existe una varianza explicada de un 73% total(de 51.3% en PC1 y 21.4% en PC2). Por otro lado, los grupos tienen un cierto grado de solapamiento pero conservan cohesión. Luego, los vectores de las variables originales nos dicen lo siguiente (Ver apéndice para una mejor descripción de las variables):

Grupo 1

- Tienen diámetro mínimo de valores medios/altos.
- Tienen una magnitud absoluta menor al promedio.
- Tienen una inclinación de su órbita (con respecto a la órbita de la tierra) media/alta.
- Tienen órbitas desde distancias promedio a muy alejadas de la órbita de la tierra.
- Tienen las órbitas menores al promedio de distancia a la órbita del sol.

Grupo 2

- Tienen un diámetro mínimo en general menor a aquellos que se encuentran en el grupo 1.
- Tienen una magnitud absoluta menor al promedio pero mayor que el grupo 1.
- Tienen una inclinación de su órbita (Con respecto a la órbita de la tierra) alta/media y mayor al grupo 1.
- Tienen órbitas desde distancias promedio a muy alejadas de la órbita de la tierra pero menores al grupo 1.
- Tienen órbitas desde valores promedio a las más lejanas a la órbita del sol.

Grupo 3

- Tienen un diámetro mínimo menor al promedio.
- Aquellos con mayor magnitud absoluta.
- Tienen una inclinación de su órbita (con respecto a la órbita de la tierra) menor al promedio.
- Tienen las órbitas menores al promedio de la distancia a la órbita de la tierra.
- Tienen órbitas desde valores promedio a las más lejanas a la órbita del sol al igual que el grupo 2

A continuación visualizamos las clases Peligroso/No peligroso en el mismo Biplot y también graficamos un barplot con la media de las poblaciones por cada grupo:

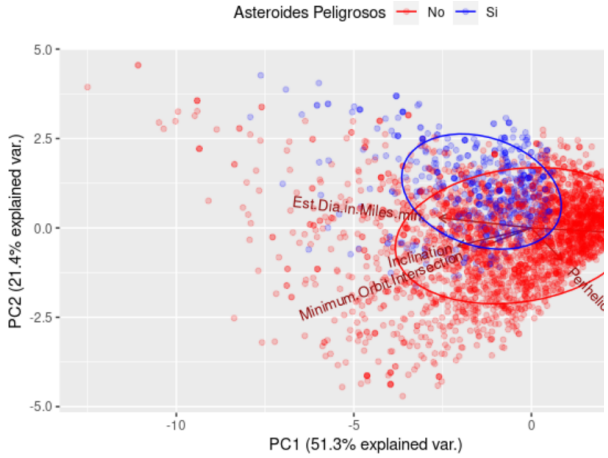

Figura 6. Se visualizan las clases reales de los individuos (peligroso/no peligroso) en un biplot.

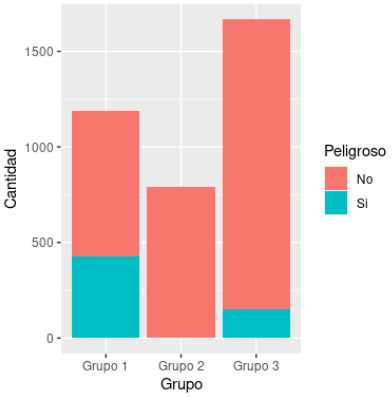

Figura 7. Se visualiza la media por población en cada grupo.

Si comparamos las figuras 1, 5, 6 y 7 podemos decir que:

- El grupo 1 es el que tiene mayor número de asteroides peligrosos seguido del grupo 2.
- El grupo 2 y 3 contiene el mayor número de asteroides no peligrosos.
- Por otro lado, el modelo parece captar las clases peligroso y no peligroso en dos grupos 1 y grupo 2+3.
- Dado que las variables **Minimum.Orbit.Intersection** y **Absolute.Magnitude** son las que mejor discriminan las clases (figura 1), podríamos decir que a valores cercanos o menores que el promedio(figura 5, 6) aumentan el riesgo de colisión.
- Dentro de los Peligroso, el grupo 2 y 3 se diferencian:
    - En mayor medida por la inclinación de su órbita.
    - En una medida media por su diámetro mínimo (**Est.Dia.in.Miles.min**) y **Minimum.Orbit.Intersection.**
    - Muy poco en su distancia a la órbita del sol (**Perihelio Distance).**

Apéndice

Descripción de variables seleccionadas

- **Absolute Magnitude**: La magnitud absoluta de un asteroide es la magnitud visual(Tamaño) que registraría un observador si el asteroide se colocará a 1 unidad astronómica (UA) de distancia y a 1 UA del Sol y en un ángulo de fase cero.
- **Est.Dia.in.Miles.min**: Diámetro mínimo del asteroide en millas.
- **Perihelion Distance**: Es el punto más cercano de la órbita del asteroide al sol.
- **Minimum.Orbit.Intersection**: Es una medida que permite evaluar el riesgo de colisión de dos cuerpos celestes. Está definida como la distancia mínima entre los puntos próximos de las órbitas de ambos cuerpos celestes.
- **Inclination:** Es el ángulo de inclusión de la órbita del asteroide con respecto al plano formado por la órbita de la tierra.
- **Hazardous**: Determina si el asteroide es peligroso o no.