


Universidad Nacional Autónoma de México

Facultad de Estudios Superiores Acatlán

Modelos de clasificación para predecir la peligrosidad de asteroides

Proyecto Final del Módulo II - Diplomado en Ciencia de Datos

Autor:
Juan Manuel Zapata López

Abril 2021

Índice

1. Introducción

Una de las mayores fascinaciones del ser humano, a lo largo de su historia, ha sido el espacio exterior. Desde antiguas civilizaciones ha habido investigaciones científicas para tratar de entender que hay más allá de nuestro planeta. Y los avances han sido significativos. Hace poco más de 200 años, un científico de la Royal Society de Londres descubrió un nuevo tipo de cuerpo espacial, al que llamó asteroide.

Los asteroides son objetos rocosos que orbitan alrededor del sol, pero con un tamaño considerablemente menor al de un planeta. Sabemos que hay una gran cantidad de asteroides en nuestro sistema solar y la mayoría se encuentra en una región denominada cinturón de asteroides, comprendida entre las órbitas de los planetas Marte y Júpiter. Algunos otros, se encuentran en las órbitas de los planetas y siguen el mismo recorrido que el planeta alrededor del sol.

Algunas de las características de estos objetos son las siguientes: no son objetos redondos como los planetas (de hecho tienen formas muy irregulares), algunos asteroides tienen cientos de kilómetros de diámetro, pero la mayoría son del tamaño de un puño humano. Los materiales de los que están formados suelen ser arcilla, níquel y hierro.

Actualmente, la NASA se encarga de monitorear estos objetos con el fin de entender mejor nuestro universo, pero también de prevenir que alguno de ellos se convierta en un peligro potecial para los habitantes de la Tierra. Alrededor de nuestro planeta, podemos encontrar diversos cráteres que testifican como hace miles de años hubo asteroides que impactaron y pudieron causar grandes repercusiones. Por ejemplo, se cree que uno de estos eventos fue el causante de la extinción de los dinosaurios.

1.1. Descripción del dataset

En este contexto, surge la curiosidad de investigar si existe algún tipo de medida para la peligrosidad de un asteroide. A través de la página https://www.kaggle.com/shrutimehta/nasa-asteroids-classification se enontró una tabla con características de poco más de 4 mil asteroides distintos, que ha sido recopilada por la NASA en distintas fechas y que al final ha clasificado con base en el riesgo que representan.

1.2. Objetivo

El objetivo del presente texto, es aplicar técnicas de limpieza de datos, detección de outliers y aplicación algunas otras opciones para hacer un análisis exploratorio del dataset y crear una tabla analítica de datos, con la cuál probar distintos métodos de aprendizaje supervisado para clasificación y finalmente elegir un modelo que nos ayude a entender como la NASA realiza la clasificación de los asteroides por su peligrosidad.

2. Data Set

2.1. Descripción de la información

En el dataset que se seleccionó para analizar, encontramos inicialmente un total de 4,687 registros con 40 variables que se describen a continuación.

Variable	Descripción
Neo Reference ID	Id del asteroide
Name	Nombre del asteroide
Absolute Magnitude	Magnitud visual si el asteroide se colocara a 1 AU del Sol
Est Dia in KM(min)	Diámetro mínimo estimado en km
Est Dia in KM(max)	Diámetro máximo estimado en km
Est Dia in M(min)	Diámetro mínimo estimado en metros
Est Dia in M(max)	Diámetro máximo estimado en metros
Est Dia in Miles(min)	Diámetro mínimo estimado en millas
Est Dia in Miles(max)	Diámetro máximo estimado en millas
Est Dia in Feet(min)	Diámetro mínimo estimado en pies
Est Dia in Feet(max)	Diámetro máximo estimado en pies
Close Approach Date	Fecha de un acercamiento a menos de una distancia lunar de la Tierra
Epoch Date Close Approach	Fecha de un acercamiento en epoch
Relative Velocity km per sec	Velocidad relativa en km por segundo
Relative Velocity km per hr	Velocidad relativa en km por hora
Miles per hour	Velocidad relativa en millas por hora
Miss Dist.(Astronomical)	Distancia de pérdida en unidades astronómicas
Miss Dist.(lunar)	Distancia de pérdida en distancia lunar
Miss Dist.(kilometers)	Distancia de pérdida en km
Miss Dist.(miles)	Distancia de pérdida en millas
Orbiting Body	Planeta alrededor del cuál el asteroide orbita
Orbit ID	Id de la órbita
Orbit Determination Date	Fecha en que se determinó la órbita
Orbit Uncertainity	Métrica que cuantifica diversos parámetros de la órbita
Minimum Orbit Intersection	Distancia entre los puntos más cercanos de la órbita con otra
Jupiter Tisserand Invariant	Parámetro Tisserand del asteroide
Epoch Osculation	El instante en el que se especifican los vectores de posición y velocidad
Eccentricity	Excentricidad de la órbita
Semi Major Axis	Valor del eje mayor de la órbita
Inclination	Inclinación de la órbita
Asc Node Longitude	Elemento orbital
Orbital Period	Tiempo que demora el asteroide en completar una órbita
Perihelion Distance	Distancia mínima al sol
Perihelion Arg	Argumento del perihelio
Aphelion Dist	Distancia de Apheliom
Perihelion Time	Tiempo de perihelio
Mean Anomaly	la fracción del período de la órbita desde que el asteroide pasó la periapsis
Mean Motion	Velocidad angular que se requiere para completar una órbita
Equinox	Equinoccio del asteroide
Hazardous	Variable objetivo, denota si un asteroide es peligroso o no

Cuadro 1: Descripción de variables

3. Calidad de datos

3.1. Duplicidad

Se buscaron líneas duplicadas en la tabla y también registros duplicados por ID, pero no se encontraron.

3.2. Verificación del tipo de dato

Se obtuvo el tipo de dato de cada variable, siendo la mayoría int64 o float64. Para las variables con tipo object se realizó una examinación para verificar si eran de tipo texto, fecha o contenían errores. Se encontraron dos variables de tipo fecha que fueron transformadas a datetime. Dos variables eran de tipo de texto(Orbiting Body y Equinox), pero contenían el mismo valor en todos los registros y fueron eliminadas.

3.3. Completitud

La completitud es la cantidad de registros que no tienen missings expresada en porcentaje. Para esta tabla, se calculó la completitud de cada variable y se encontró que no existían registros nulos.

4. Identificación y remoción de outliers

En este caso, se creó una función que detecta outliers primero por rango intercuartil, después por percentiles .5 y .95 y finalmente por z-score. Para z-score, primero se aplicó una prueba de normalidad, si era rechazada se hacía una transformación de box-cox para intentar inducir la normalidad, si ésta era aceptada se hacía z-score, en otro caso, este método no se usaba. También, se decidió aplicar esta meodología sólo a las varaibles que quedaron después de reducir dimensiones (este punto se analizará más adelante).

Estos son los resultados:

Para la columna Relative Velocity km per sec se tiene: Número de outliers por IQR: 101. Número de outliers por percentiles .5 y .95: 470. Los datos no distribuyen normal, se hará una transformación de boxcox. los datos no se pudieron tranformar a una normal, no se puede aplicar z-score. Hay 101 outliers comúnes = 0.0215 %

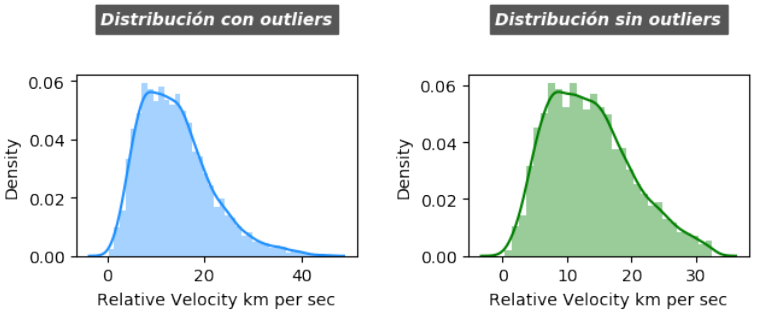


Figura 1: Variable Relative Velocity km per sec con y sin outliers

Para la columna Minimum Orbit Intersection se tiene: Número de outliers por IQR: 198. Número de outliers por percentiles .5 y .95: 459. Los datos no distribuyen normal, se hará una transformación de boxcox. los datos no se pudieron tranformar a una normal, no se puede aplicar z-score. Hay 198 outliers comúnes = 0.0432 %

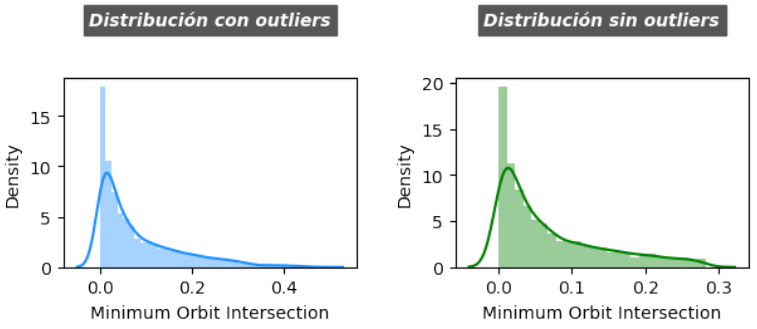


Figura 2: Variable Minimum Orbit Intersection con y sin outliers

Para la columna Eccentricity se tiene: Número de outliers por IQR: 2. Número de outliers por percentiles .5 y .95: 440. Los datos no distribuyen normal, se hará una transformación de boxcox. los datos no se pudieron tranformar a una normal, no se puede aplicar z-score. Hay 2 outliers comúnes = 0.0005 %

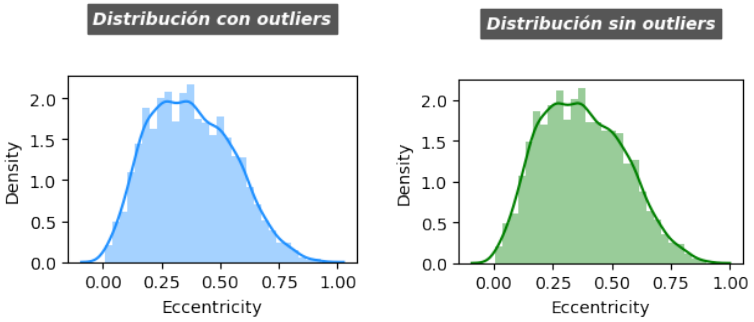


Figura 3: Variable Eccentricity con y sin outliers

Para la columna Inclination se tiene: Número de outliers por IQR: 93. Número de outliers por percentiles .5 y .95: 439. Los datos no distribuyen normal, se hará una transformación de boxcox. los datos no se pudieron tranformar a una normal, no se puede aplicar z-score. Hay 93 outliers comúnes = 0.0212 %

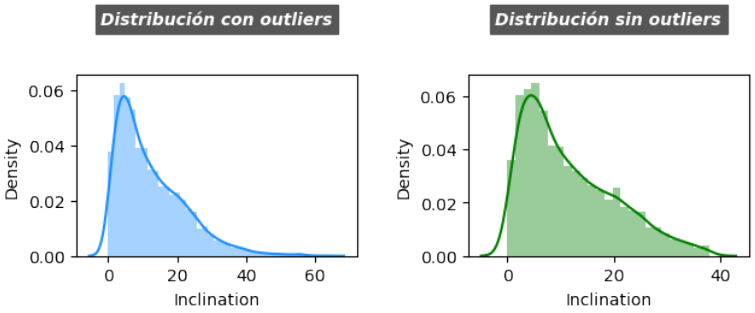


Figura 4: Variable Inclination con y sin outliers

5. Reducción de dimensiones

Antes de proseguir, es importante mencionar que dado que no habían missings en ninguna variable no hubo necesidad de aplicar técnicas de imputación de valores. También, las variables que hay hasta ahora en la tabla son todas de tipo numérico, salvo algunos ids y fechas que no aportan al modelo y fueron eliminadas. Por ello se puede iniciar con las técnicas de reducción de dimensiones.

5.1. Filtro de baja varianza

Se obtuvo la varianza de cada variable, no se encontró alguna varianza baja, por lo cual ninguna variable fue eliminada en este filtro.

5.2. Filtro de alta correlación

Para iniciar, se graficó a la matriz de correlaciones de las variables explicativas, para tener una vista general de como se comportan las variables entre sí, ésta fue calculada con el método de pearson.

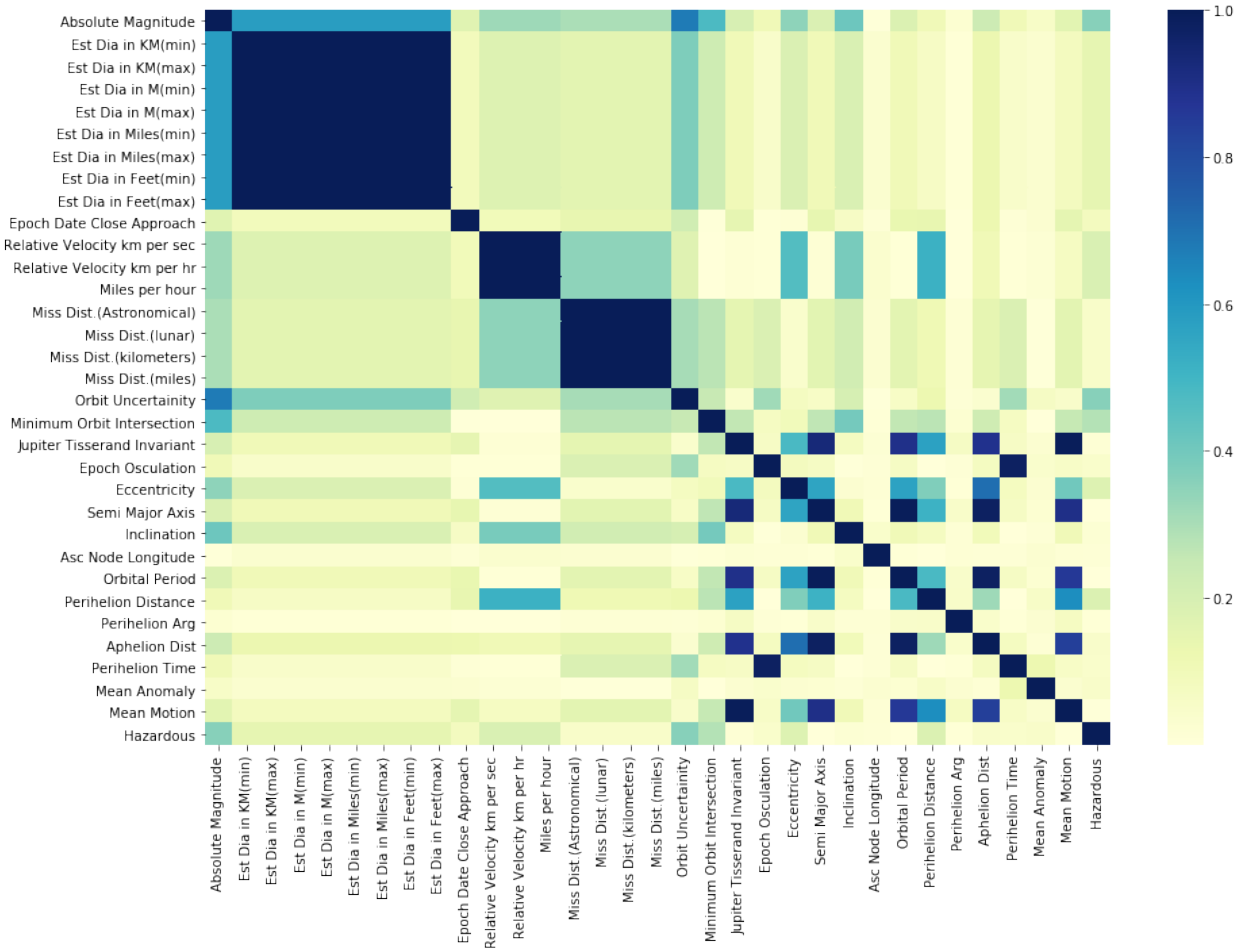


Figura 5: Matriz de correlaciones antes de eliminar variables altamente colacionadas

Se aprecia que hay variables con alta correlación, lo cuál tiene mucho sentido, pues algunas de ellas cuantifican lo mismo, solo se presentan en diferente escala (metros, kilómetros, millas). Se removieron todas las variables cuya correlación fuera mayor a 0.8, siendo las siguientes (se incluye a la variable objetivo), las sobrevivientes:

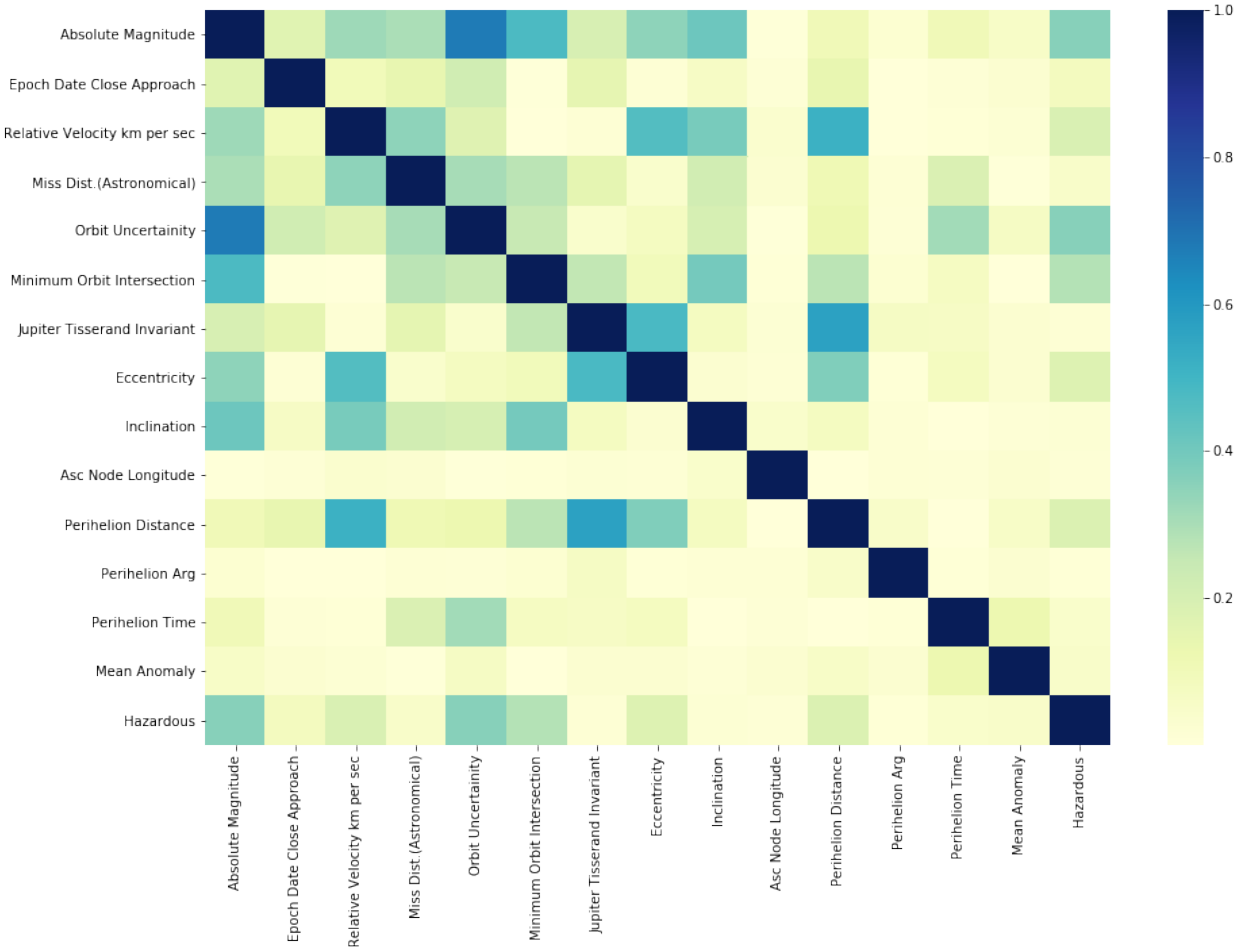


Figura 6: Matriz de correlaciones después de eliminar variables altamente colacionadas

5.3. Filtro de baja correlación con la variable objetivo

Para las variables que aún tenemos en la tabla, se calculó su correlación de pearson con la variable objetivo, se tienen los siguiente resultados:

	Hazardous
Hazardous	1.000000
Absolute Magnitude	0.383620
Orbit Uncertainity	0.352155
Minimum Orbit Intersection	0.259885
Perihelion Distance	0.191258
Relative Velocity km per sec	0.188888
Eccentricity	0.172256
Epoch Date Close Approach	0.083097
Miss Dist.(Astronomical)	0.054827
Mean Anomaly	0.050921
Perihelion Time	0.034033
Jupiter Tisserand Invariant	0.024411
Asc Node Longitude	0.013849
Inclination	0.012084
Perihelion Arg	0.011618

Figura 7: Correlaciones con la variable objetivo

Se observa que en realidad son muchas las variables que presentan baja correlación con la variable target, se decide eliminar aquellas cuya correlación sea la menor.

5.4. Multicolinealidad

Para ver si existen variables que puedan inducir a problemas de multicolinealidad, se calculó el factor de inflación de la varianza:

variables	VIF
Absolute Magnitude	33.30
Perihelion Distance	18.57
Relative Velocity km per sec	12.86
Eccentricity	8.18
Miss Dist.(Astronomical)	5.55
Inclination	4.61
Orbit Uncertainity	3.62
Minimum Orbit Intersection	3.54

Figura 8: calculo del VIF por variable

Aunque se observa que la variable Absolute Magnitude presenta un VIF alto, se decidió no eliminarla porque es la que presenta mayor correlación con la variable objetivo. Ninguna variable se eliminó en este filtro.

6. Análisis exploratorio

Con el objetivo de entender el contenido de la TAD, se analizó de forma gráfica cada una de las variables. Se inició haciendo boxplots de cada variable explicativa dependiendo de si los meteoritos son peligrosos o no lo son.

En estas primeras 4 gráficas, se observa que la distancia de pérdida y la velocidad relativa son muy similares para meteoritos peligrosos y no peligrosos. Mientras que la magnitud absoluta suele tomar valores entre 15 y 23 (con una mediana de 20) para asteroides peligrosos y toma valores en un rango más amplio , desde 10 hasta 35, pero con una mediana de 23 para asteroides no peligrosos. Para asteroides peligrosos, la mayoría toma valores entre 0 y 2 para la incertidumbre de la órbita, mientras que asteroides no peligrosos suelen tener valores más altos, llegando hasta un máximo de 10.

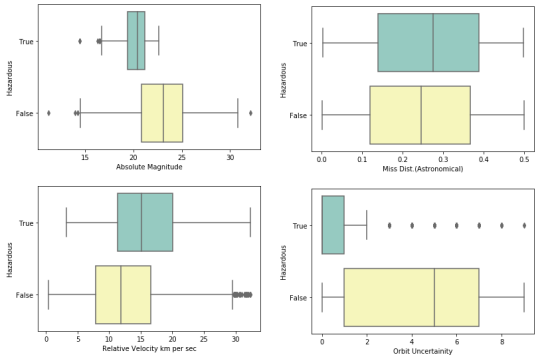


Figura 9: Boxplots de las primeras 4 variables explicativas

Ahora, en estos siguientes boxplots se observa que la inclinación, excentricidad y distancia perihelio son muy similares para ambos tipos de asteriodes. Sin embargo, la variable intersección mínima de la órbita suele tomar valores muy pequeños para asteroides peligrosos (entre 0 y 0.05), mientras que para asteroides no peligrosos toma valores en un rango más amplio (desde 0 hasta 0.3) y esto tiene mucho sentido, pues esta variable cuantifica que tan cerca está una órbita de otra y con ello mide el riesgo de impacto entre un asteroide y otro cuerpo.

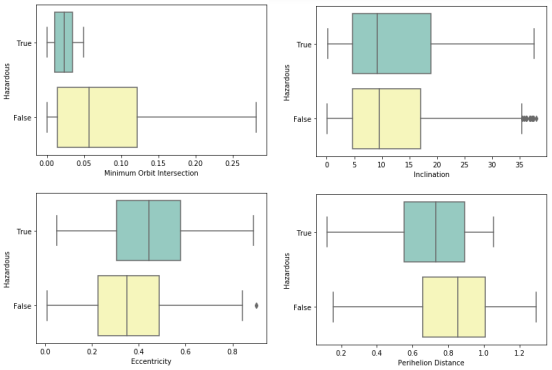


Figura 10: Boxplots de las otras variables explicativas

Si analizamos la variable ïncertidumbre de la órbita", podemos notar que toma valores entre cero y 9, pero la mayoría de asteroides tienen el valor de cero.

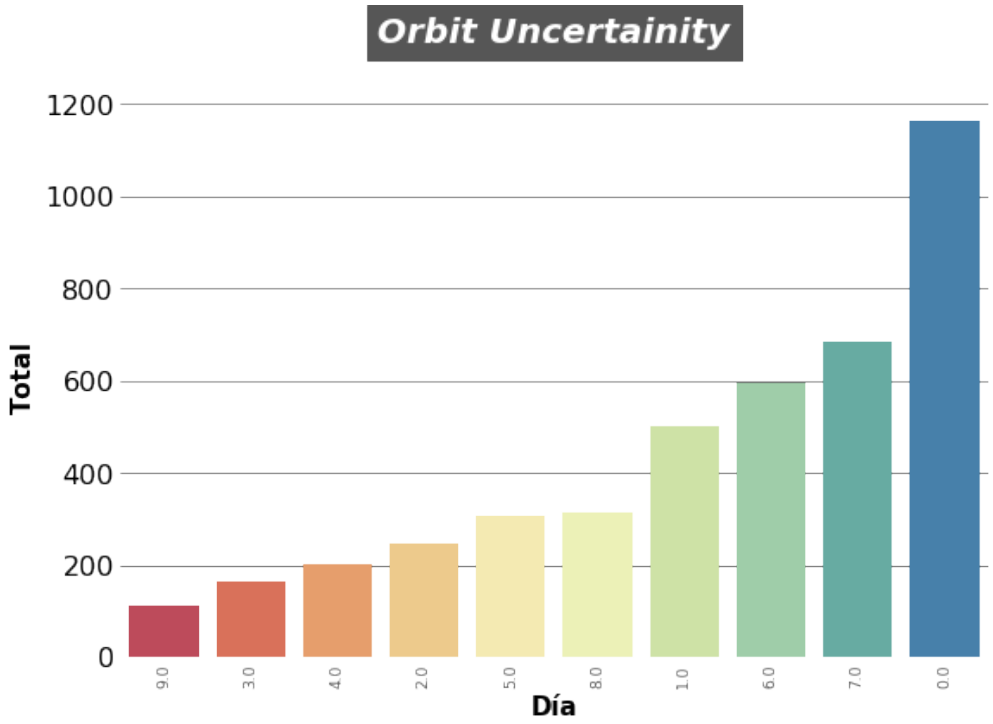


Figura 11: Gráfica de barras de incertidumbre de la órbita

Al hacer una gráfica de dispersión de la magnitud absoluta vs la intersección mínima de órbita, se percibe una pequeña correlación negativa entre ambas variables, pero lo más importante es que al hacer la distinción por tipo de asteroide, se nota que las clases son separables casi de forma perfecta, salvo algunos puntos que se enciman.

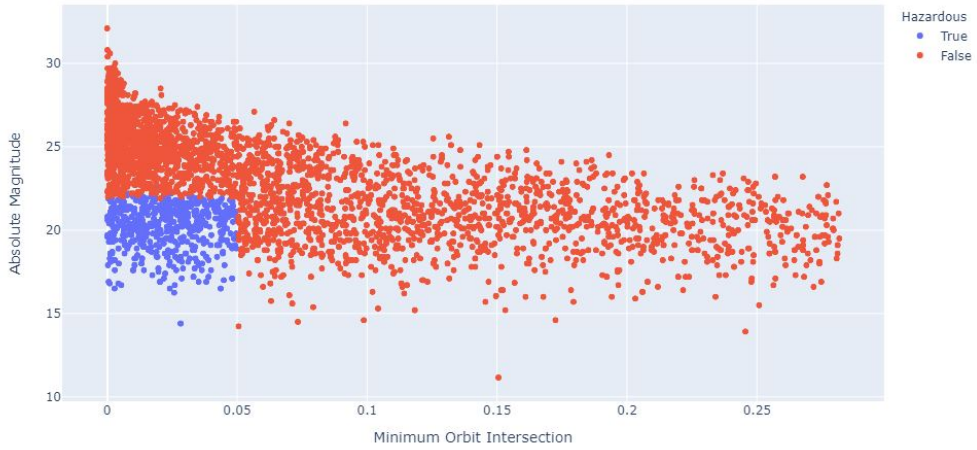


Figura 12: Gráfica de dispersión

Si añadimos una tercera dimensión a la gráfica anterior (con la variable incertidumbre de la órbita), también se aprecia la separación de clases, y notamos que la mayoría de asteroides peligrosos se concentra en magnitudes absolutas e intersecciones mínimas de órbita menores, mientras que se encuentran bien distribuidos por incertidumbre de la órbita.

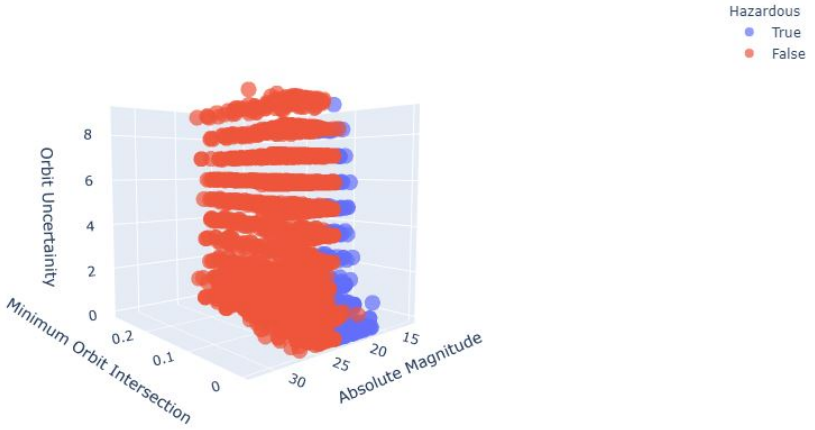


Figura 13: Gráfica de dispersión en 3 dimensiones

Finalmente, al hacer boxplots con dos variables al mismo tiempo, sólo reafirmamos lo visto anteriormente, asteroides peligrosos suelen tener menor magnitud absoluta, sin importar la incertidumbre de la órbita.

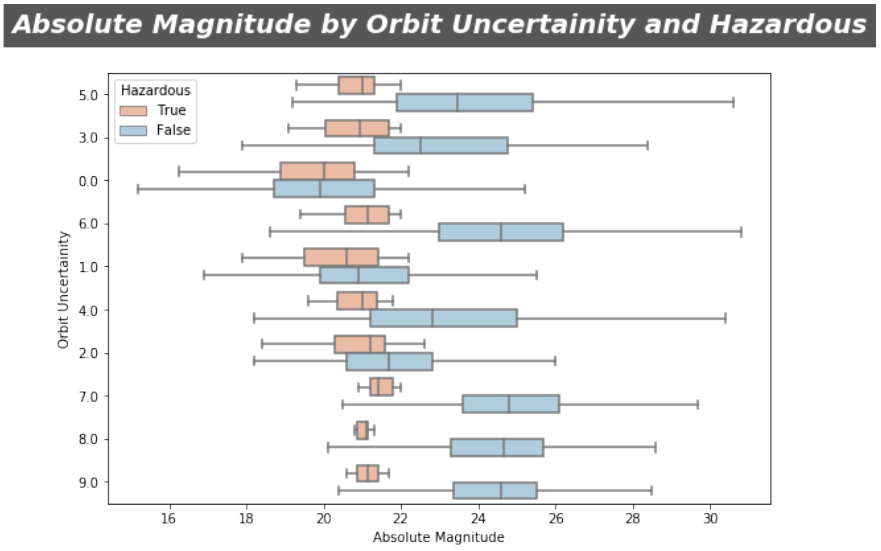


Figura 14: Boxplot multivariable

7. Modelos supervisados para clasificación

Antes de comenzar con esta sección, es bueno resaltar que se probaron distintos modelos con balanceo y sin balanceo de clases, en el presente documento sólo se muestran algunos. Para monitorear si el modelo es bueno, se crearon funciones para visualizar sus métricas, matriz de confusión, gráfica ROC y estabilidad. Además, para encontrar los mejores hiperparámetros se utilizó una búsqueda aleatoria.

7.1. División en train y test

En este punto se tiene una tabla con 4,293 registros, 8 variables explicativas y una variable objetivo. Los porcentajes de división que se usaron para el conjunto de entrenamiento y prueba son de 70 % y 30 % respectivamente.

7.2. Modelos sin balancear

Al verificar el balanceo tanto en el conjunto de entrenamiento como en el de prueba se obtuvieron los siguientes resultados:

```
y_train.value_counts(1)
```
executed in 12ms, finished 15:21:11 2021-04-25
```
0.00   0.84
1.00   0.16
Name: Hazardous, dtype: float64
```
```
y_test.value_counts(1)
```
executed in 12ms, finished 15:21:11 2021-04-25
```
0.00   0.84
1.00   0.16
Name: Hazardous, dtype: float64
```

Figura 15: Balanceo de clases

La proporción de clases se mantiene en ambos conjuntos y la clase positiva (asteroide peligroso) presenta un porcentaje bajo. Sin embargo, se van a probar modelos con estos porcentajes para contrastarlos con modelos en los que previamente se aplicaron técnicas de balanceo.

7.2.1. Support Vector Machine

En este modelo los mejores hiperparámetros fueron un kernel lineal de grado 5.

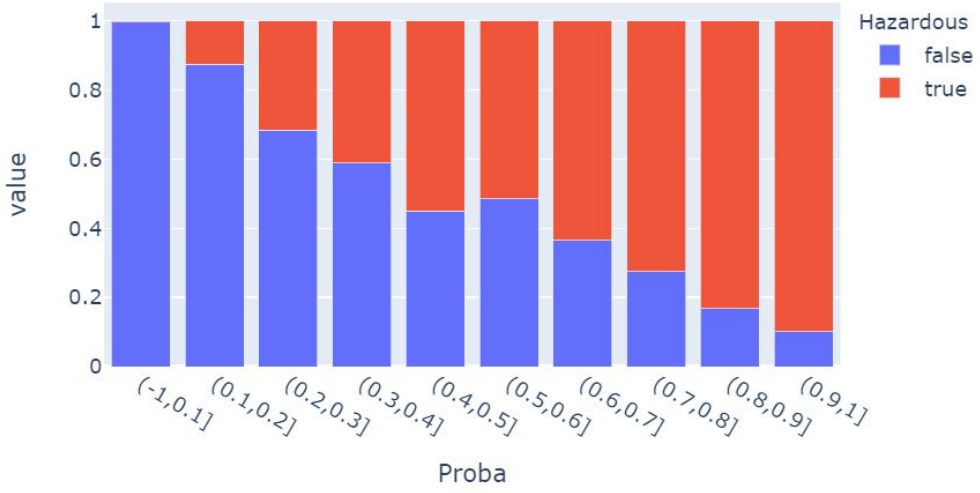
Las metricas en train son las siguientes:

	Accuracy	Precision	Recall	F1	TPR	FPR	AUC
0	0.91	0.84	0.56	0.67	0.56	0.02	0.97

Las metricas en test son las siguientes:

	Accuracy	Precision	Recall	F1	TPR	FPR	AUC
0	0.92	0.86	0.61	0.72	0.61	0.02	0.97

Y la estabilidad del modelo se ve así:



El modelo es muy bueno, pues se obtuvieron buenas métricas en el entrenamiento y aún un poco mejores en la prueba. Sin embargo el recall y la tasa de verdaderos positivos están un poco bajos. La estabilidad del modelo es buena, pues e percibe un incremento gradual de la clase positiva conforme la probabilidad aumenta.

7.2.2. Árbol de decisión

En este modelo los mejores hiperparámetros fueron un criterio de divisibilidad de Gini y una profundidad máxima de 18.

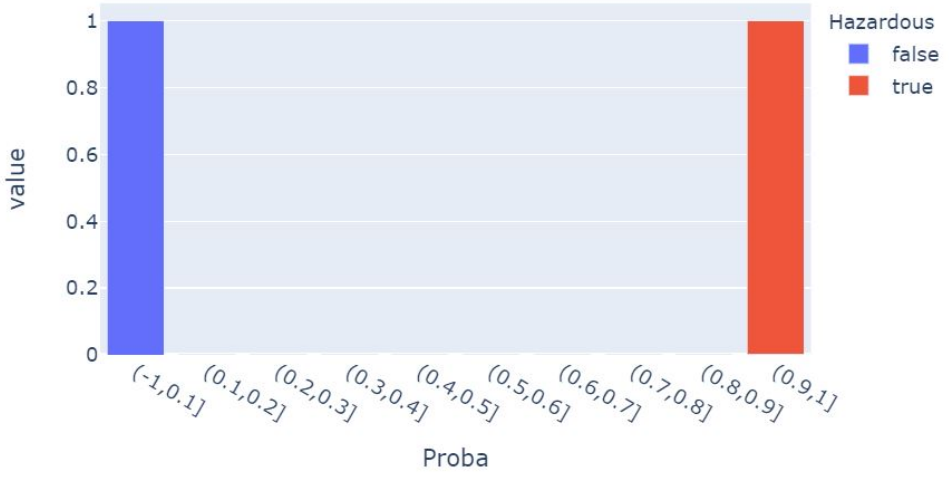
Las metricas en train son las siguientes:

	Accuracy	Precision	Recall	F1	TPR	FPR	AUC
0	1.00	1.00	1.00	1.00	1.00	0.00	1.00

Las metricas en test son las siguientes:

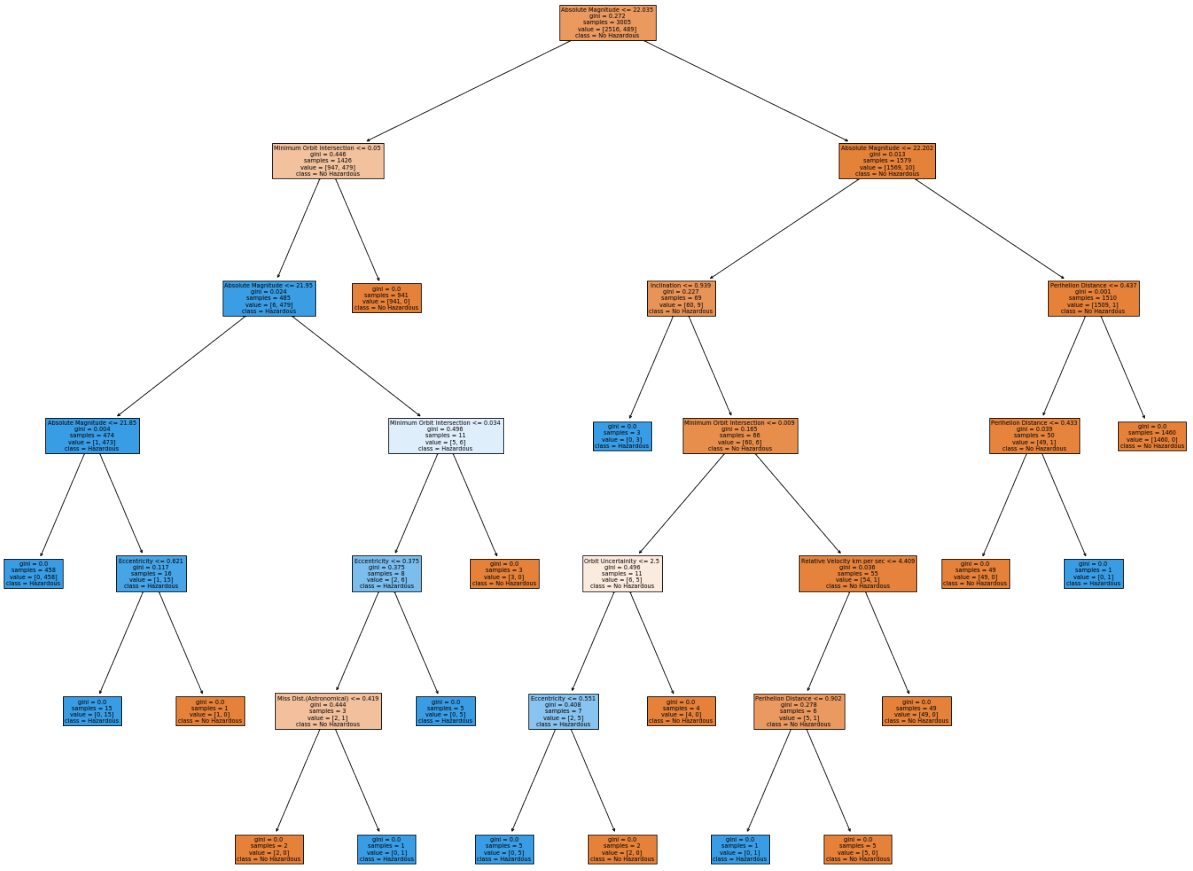
	Accuracy	Precision	Recall	F1	TPR	FPR	AUC
0	1.00	0.99	0.99	0.99	0.99	0.00	0.99

Y la estabilidad del modelo se ve así:



El modelo está sobreajustado, pues todas las métricas en el entrenamiento son de 1. Esto quizás se pueda evitar poniendo más hiperparámetros, como un mínimo de registros en cada hoja. La estabilidad no se aprecia, pues el calculo de probabilidades en árboles no guarda mucho sentido con el concepto, todos los registros clasificados como no peligrosos tienen probabilidad menor a .1 y todos los registros clasificados como peligrosos tienen probabilidad mayor a 0.9.

Lo bueno de este modelo, es que tiene interpretabilidad, sólo con el objeto de conocer la forma del árbol, este se graficó (las reglas no se alcanzan a percibir).


Absolute Magnitude <= 22.035
gini = 0.212
samples = 3003
value = [2516, 489]
class = No Hazardous
Minimum Orbit Intersection <= 0.05
gini = 0.446
samples = 1426
value = [947, 479]
class = No Hazardous
Absolute Magnitude <= 22.202
gini = 0.053
samples = 1579
value = [1569, 10]
class = No Hazardous
Absolute Magnitude <= 21.95
gini = 0.024
samples = 485
value = [6, 479]
class = Hazardous
gini = 0.0
samples = 941
value = [941, 0]
class = No Hazardous
Inclination <= 0.939
gini = 0.227
samples = 69
value = [60, 9]
class = No Hazardous
Perihelion Distance <= 0.437
gini = 0.001
samples = 1510
value = [1509, 1]
class = No Hazardous
Absolute Magnitude <= 21.85
gini = 0.004
samples = 474
value = [1, 473]
class = Hazardous
Minimum Orbit Intersection <= 0.034
gini = 0.496
samples = 11
value = [5, 6]
class = Hazardous
gini = 0.0
samples = 3
value = [0, 3]
class = Hazardous
Minimum Orbit Intersection <= 0.009
gini = 0.165
samples = 66
value = [60, 6]
class = No Hazardous
Perihelion Distance <= 0.433
gini = 0.039
samples = 50
value = [49, 1]
class = No Hazardous
gini = 0.0
samples = 1460
value = [1460, 0]
class = No Hazardous
gini = 0.0
samples = 458
value = [0, 458]
class = Hazardous
Eccentricity <= 0.621
gini = 0.117
samples = 16
value = [1, 15]
class = Hazardous
Eccentricity <= 0.375
gini = 0.375
samples = 8
value = [2, 6]
class = Hazardous
gini = 0.0
samples = 3
value = [3, 0]
class = No Hazardous
Orbit Uncertainty <= 2.5
gini = 0.496
samples = 11
value = [6, 5]
class = No Hazardous
Relative Velocity km per sec <= 4.405
gini = 0.036
samples = 55
value = [54, 1]
class = No Hazardous
gini = 0.0
samples = 49
value = [49, 0]
class = No Hazardous
gini = 0.0
samples = 1
value = [0, 1]
class = Hazardous
gini = 0.0
samples = 15
value = [0, 15]
class = Hazardous
gini = 0.0
samples = 1
value = [1, 0]
class = No Hazardous
Miss Dist.(Astronomical) <= 0.419
gini = 0.444
samples = 3
value = [2, 1]
class = No Hazardous
gini = 0.0
samples = 5
value = [5, 0]
class = Hazardous
Eccentricity <= 0.555
gini = 0.408
samples = 7
value = [2, 5]
class = Hazardous
gini = 0.0
samples = 4
value = [4, 0]
class = No Hazardous
Perihelion Distance <= 0.902
gini = 0.278
samples = 6
value = [5, 1]
class = No Hazardous
gini = 0.0
samples = 49
value = [49, 0]
class = No Hazardous
gini = 0.0
samples = 2
value = [2, 0]
class = No Hazardous
gini = 0.0
samples = 1
value = [0, 1]
class = Hazardous
gini = 0.0
samples = 5
value = [0, 5]
class = Hazardous
gini = 0.0
samples = 2
value = [2, 0]
class = No Hazardous
gini = 0.0
samples = 1
value = [0, 1]
class = Hazardous
gini = 0.0
samples = 5
value = [5, 0]
class = No Hazardous

7.2.3. Random Forest

En este modelo los mejores hiperparámetros fueron los siguientes: oob_score: False, n_estimators: 290, min_samples_leaf: 0.01, max_features: log2, max_depth: 7, criterion: gini, bootstrap: True. En esta ocasión se colocó en la gradilla que al menos haya el 1 % de registros en los nodos hoja para evitar el sobreajuste.

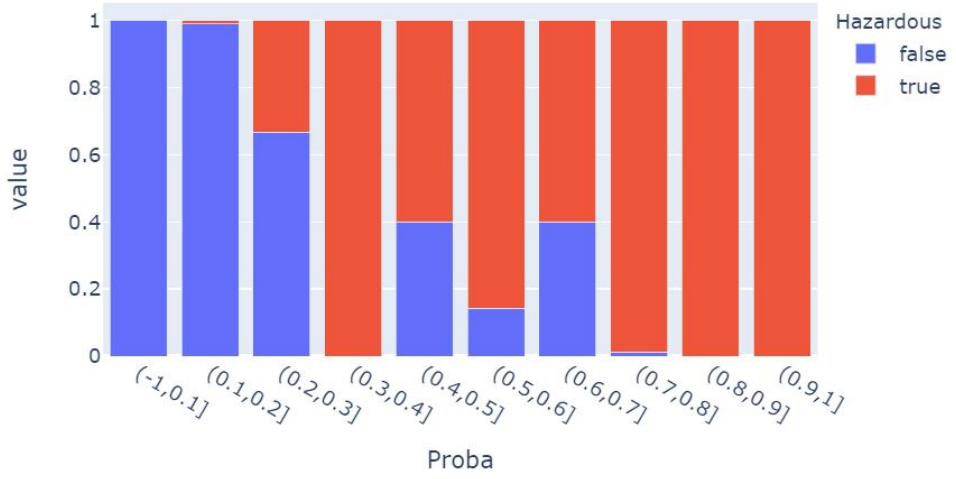
Las métricas en train son las siguientes:

	Accuracy	Precision	Recall	F1	TPR	FPR	AUC
0	0.99	0.99	0.98	0.98	0.98	0.00	1.00

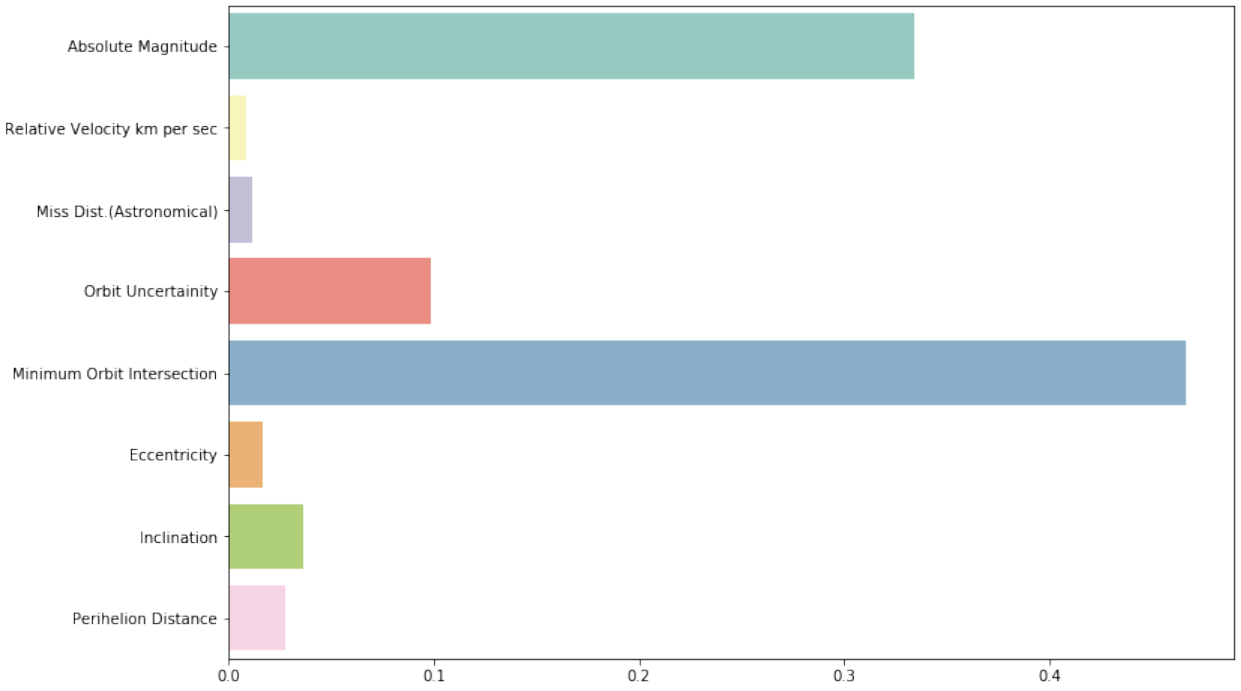
Las métricas en test son las siguientes:

Accuracy	Precision	Recall	F1	TPR	FPR	AUC
1.00	0.99	0.99	0.99	0.99	0.00	1.00

Y la estabilidad del modelo se ve así:



El modelo también es muy bueno, sus métricas son excelentes tanto en el entrenamiento como en la prueba. Sin embargo, la estabilidad del modelo no es la mejor, pues en el cuarto intervalo de probabilidad sólo hay asteroides de la clase peligrosos. Además, este modelo tiene la ventaja de asignar a cada variable su importancia al predecir, para este caso, tenemos la siguiente gráfica:



Podemos notar como las variables más influyentes son la magnitud absoluta, la intersección mínima de órbita y la incertidumbre de la órbita.

7.3. Modelos con Rebalanceo

A pesar de tener buenas métricas con los modelos sin balancear, se decidió rebalancear las clases de la variable objetivo con oversampling, es decir, se aplicó remuestreo con reemplazo a la clase minoritaria del conjunto de entrenamiento, hasta hacerla del mismo tamaño que la clase mayoritaria:

```
y_train_ov.value_counts()
```

executed in 20ms, finished 15:22:25 2021-04-2

```
0.00    2516
1.00    2516
Name: Hazardous, dtype: int64
```

7.3.1. K vecinos más cercanos

En este caso el mejor hiperparámetro fue k = 19.

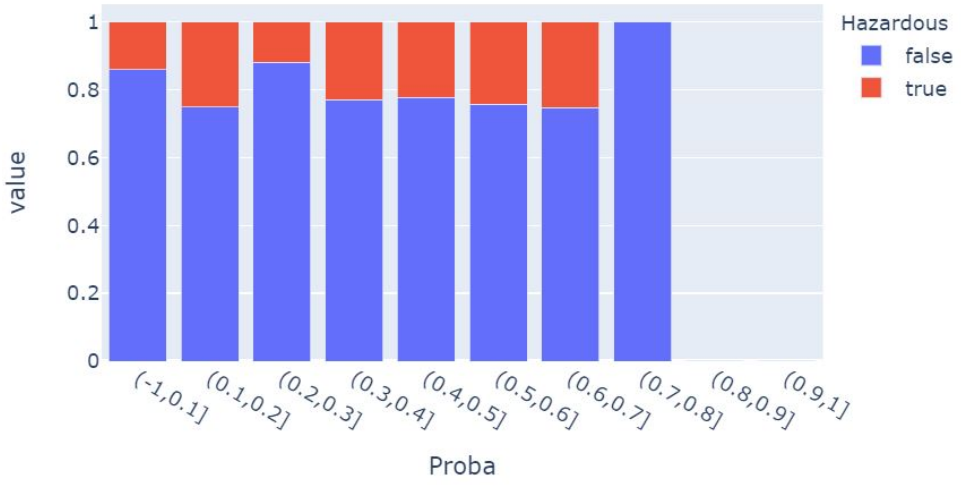
Las métricas en train son las siguientes:

	Accuracy	Precision	Recall	F1	TPR	FPR	AUC
0	0.93	0.88	0.99	0.94	0.99	0.13	0.99

Las métricas en test son las siguientes:

	Accuracy	Precision	Recall	F1	TPR	FPR	AUC
0	0.89	0.59	0.99	0.74	0.99	0.13	0.98

Y la estabilidad del modelo se ve así:



El modelo obtuvo buenas métricas en el entrenamiento, sin embargo la precisión del test está un poco baja. La estabilidad del modelo no es buena, se observan más o menos las mismas proporciones entre clases para cada nivel de probabilidad.

7.3.2. Regresión Logística

En este caso los mejores hiperparámetros son: solver = liblinear y penalidad de tipo l2.

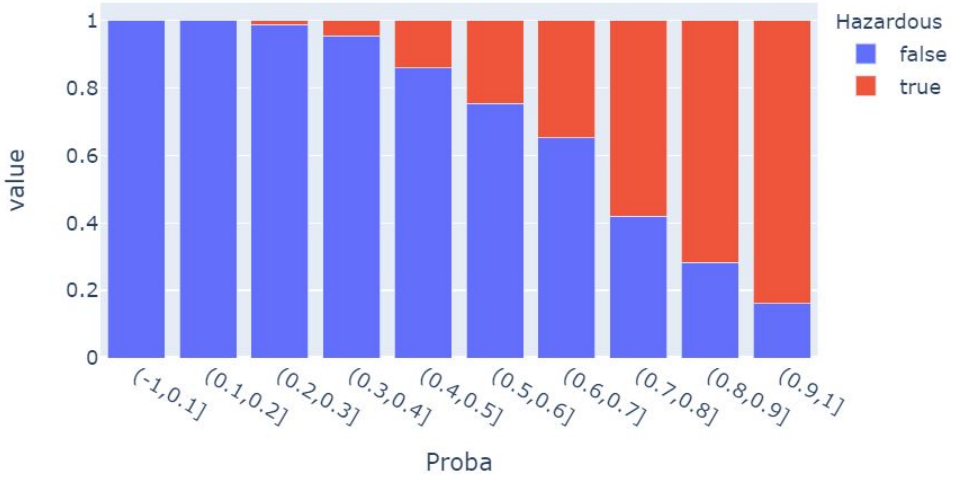
Las métricas en train son las siguientes:

	Accuracy	Precision	Recall	F1	TPR	FPR	AUC
0	0.89	0.87	0.93	0.90	0.93	0.14	0.96

Las métricas en test son las siguientes:

	Accuracy	Precision	Recall	F1	TPR	FPR	AUC
0	0.89	0.61	0.94	0.74	0.94	0.12	0.97

Y la estabilidad del modelo se ve así:



Nuevamente tenemos el caso donde el modelo obtuvo buenas métricas en el entrenamiento, sin embargo la precisión del test está un poco baja. La estabilidad del modelo es la mejor hasta ahora, pues se aprecia con claridad como al aumentar la probabilidad, también hay más registros clasificados como peligrosos. Esto puede deberse a que el modelo en sí mismo arroja una probabilidad.

7.3.3. Adaboost

En este caso el mejor hiperparámetro es una tasa de aprendizaje de 0.54.

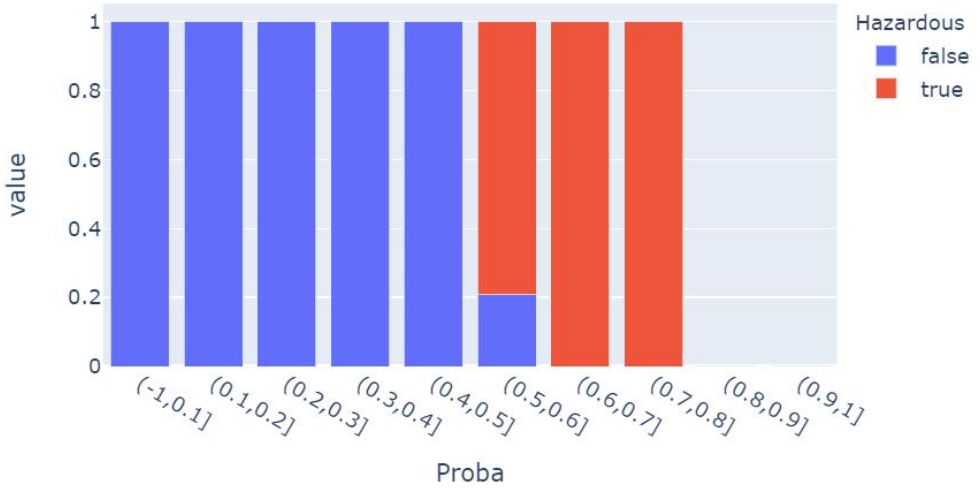
Las métricas en train son las siguientes:

	Accuracy	Precision	Recall	F1	TPR	FPR	AUC
0	1.00	1.00	1.00	1.00	1.00	0.00	1.00

Las métricas en test son las siguientes:

	Accuracy	Precision	Recall	F1	TPR	FPR	AUC
0	1.00	0.99	1.00	0.99	1.00	0.00	1.00

Y la estabilidad del modelo se ve así:



En esta ocasión, de nuevo tenemos sobreajuste pues en el entrenamiento todas las métricas son de 1. La estabilidad tampoco es la mejor, pues aunque se nota la diferencia de clases entre rangos de probabilidad, no tiene el comportamiento deseado de ver un incremento gradual.

8. Mejor Modelo

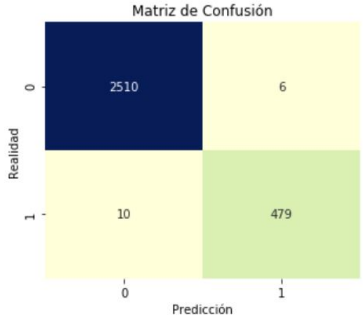
A partir de los resultados anteriores y usando el hecho de que sólo 3 variables eran las que en verdad fueron importantes (en el random forest), se decidió reducir la tabla analítica de datos a sólo esas variables explicativas y no rebalancear, puesto que no tenía un efecto de mejoría significativa.

Además, se decidió intentar con un árbol de decisión, pues se tiene la ventaja de tener interpretabilidad con las variables originales.

Con esto, los mejores hiperparámetros fueron: porcentaje mínimo de registros en nodos hoja de 1 %, criterio de gini y profundidad máxima de 26. En esta ocasión se analizaran las métricas con más detenimiento.

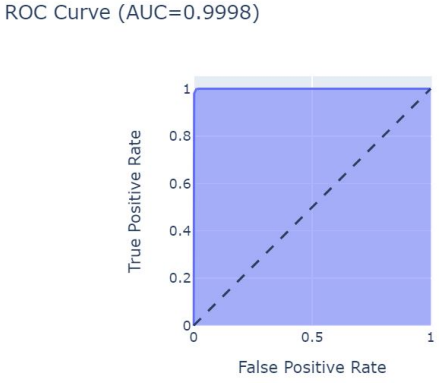
8.0.1. Métricas de entrenamiento

La matriz de confusión es la siguiente:



Son apenas 16 registros en los que el modelo no acertó.

La gráfica ROC es la siguiente:



Se aprecia como el modelo es muy bueno prediciendo en cualquier umbral de probabilidad.

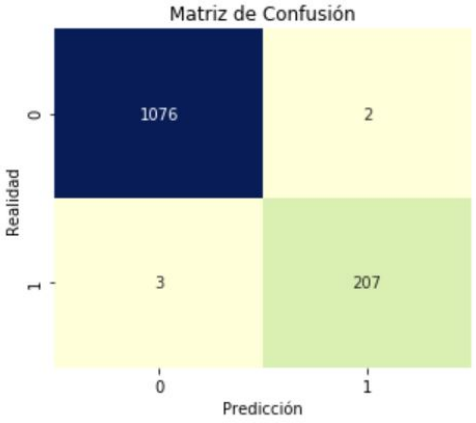
Y las métricas obtenidas son las siguientes:

	Accuracy	Precision	Recall	F1	TPR	FPR	AUC
0	0.99	0.99	0.98	0.98	0.98	0.00	1.00

Aunque son muy altas, dados los modelos anteriores podemos confiar en que no hay un sobreajuste y el modelo capta muy bien la escencia de la relación entre las variables explicativas y la objetivo.

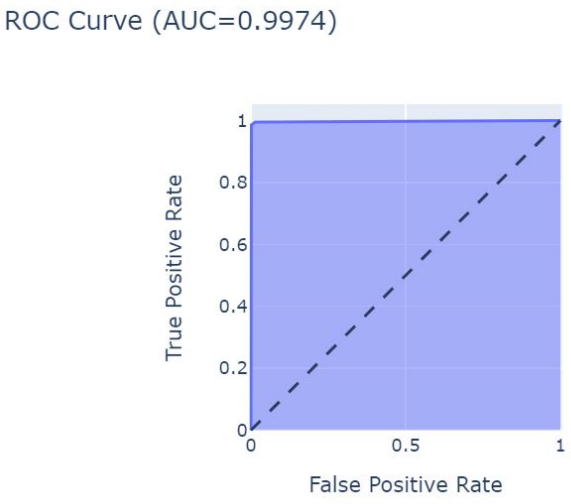
8.0.2. Métricas de prueba

La matriz de confusión es la siguiente:



Son sólo 5 registros en los que el modelo no acertó.

La gráfica ROC es la siguiente:



Igual que en el entrenamiento, se aprecia como el modelo es muy bueno prediciendo en cualquier umbral de probabilidad.

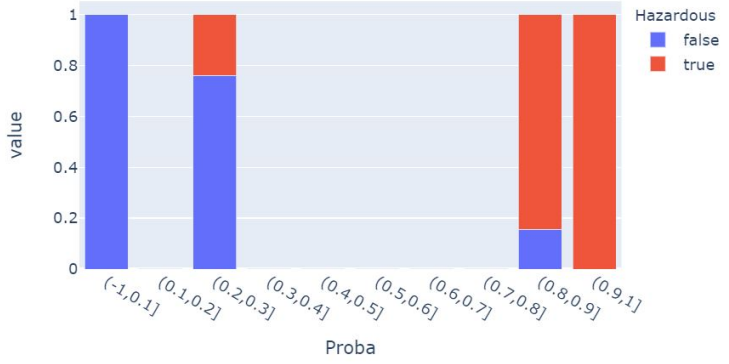
Y las métricas obtenidas son las siguientes:

	Accuracy	Precision	Recall	F1	TPR	FPR	AUC
0	1.00	0.99	0.99	0.99	0.99	0.00	1.00

Afortunadamente las métricas en el conjunto de prueba son igual de buenas que en el conjunto de entrenamiento, algunas incluso mejoraron.

8.0.3. Estabilidad del modelo

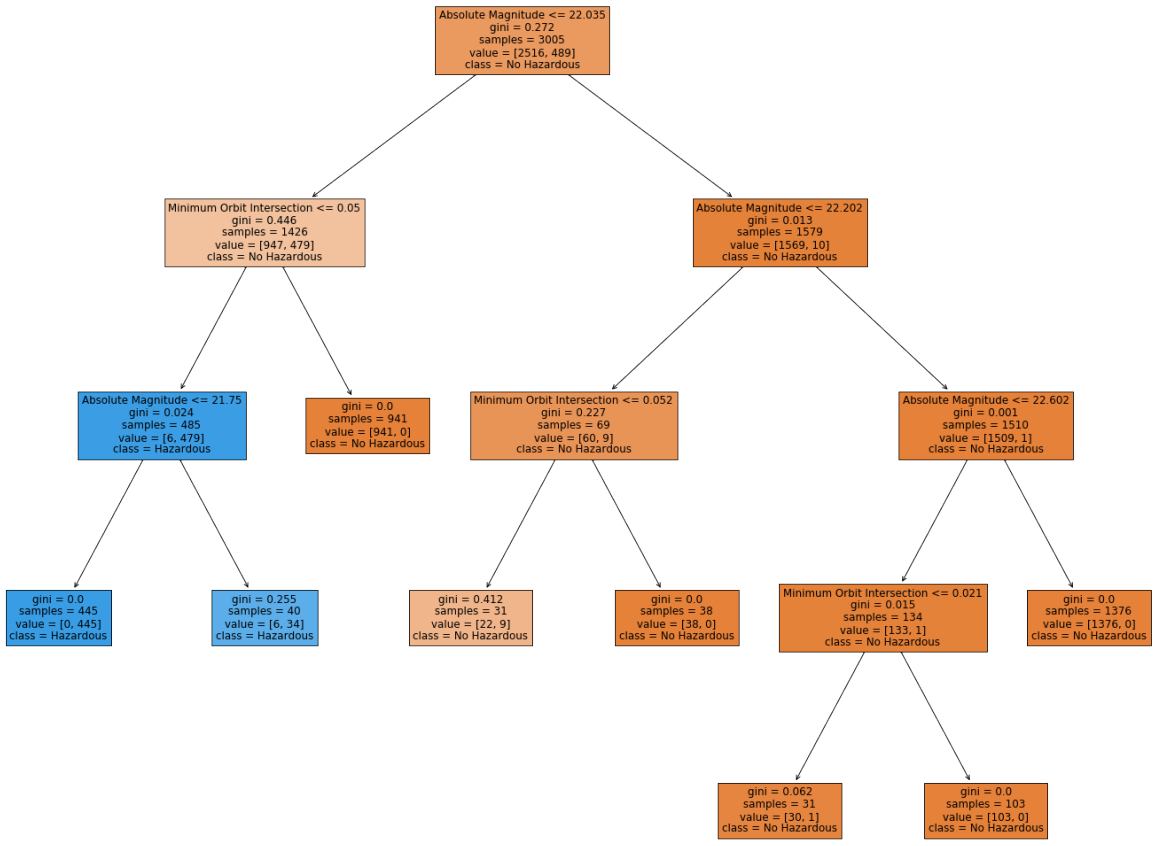
Desafortunadamente, para este modelo no tenemos la mejor estabilidad, esto se debe a que para varios intervalos de probabilidad no existen registros. Para los que sí existen, tenemos el hecho de que se aprecia como a bajos niveles de probabilidad domina la clase 'meteorito no peligroso' y conforme la probabilidad incrementa, la clase 'meteorito peligroso' gana dominio. La gráfica es la siguiente:



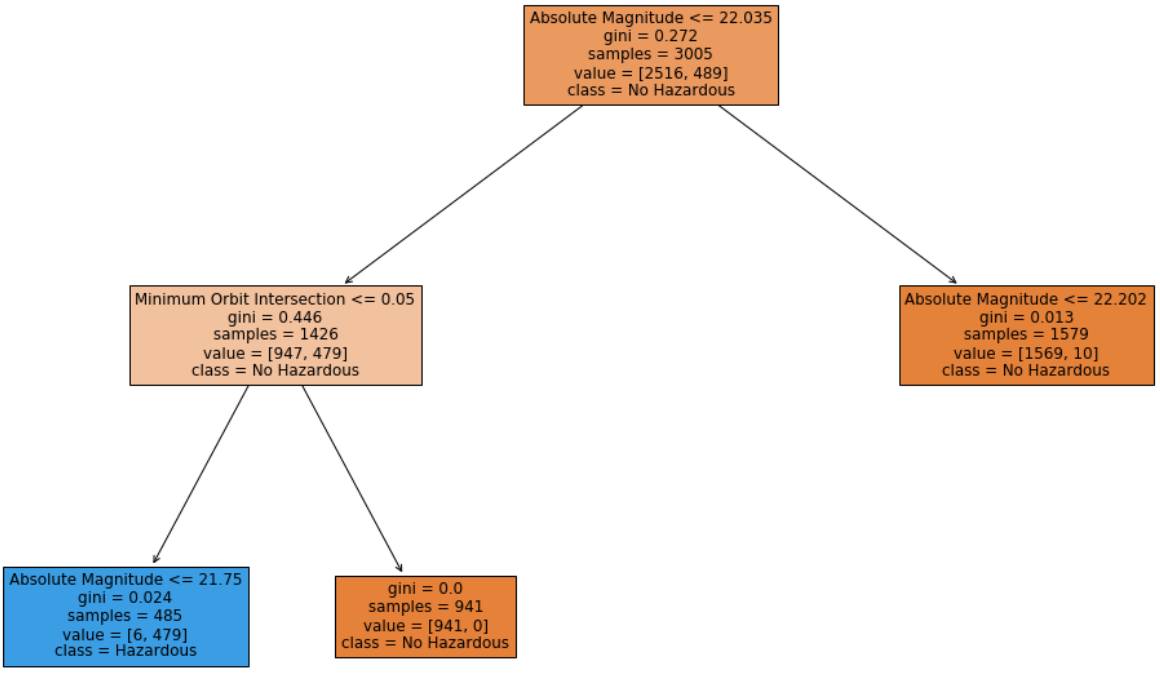
Pese a no ser el modelo con la mejor estabilidad, recordemos que ésta sirve para hacer el modelo más entendible a personas sin conocimientos técnicos y en este caso podemos omitir un poco esto, porque el modelo elegido es una serie de reglas que son fáciles de entender.

8.0.4. Interpretabilidad

Al graficar el árbol, tenemos el siguiente conjunto de reglas:



Pero el árbol puede ser podado al siguiente, que obtiene los mismos resultados pero con un conjunto de reglas más pequeño: Al graficar el árbol, tenemos el siguiente conjunto de reglas:



9. Conclusiones

Con el presente texto, conocimos un poco sobre los asteroides y su clasificación en base a los peligros que representan. Con información recopilada por la NASA se construyó un modelo de machine learning capaz de predecir de forma muy exacta e interpretable si un asteroide es peligroso o no.

La selección de variables fue acertada, pues con apenas tres (aunque el modelo al final sólo utilizó dos), se lograron resultados óptimos.