ML-Laboratory

Studenti:

- Biancini Mattia 865966
- Lorenzo Monti 869960
- Marco Gherardi 869138

Descrizione del dominio di riferimento e obiettivi dell'elaborato

Nello spazio esterno esiste un numero infinito di oggetti. Alcuni di essi sono più vicini di quanto si possa pensare. Sebbene possa sembrare che una distanza di 70.000 km non possa potenzialmente arrecare danno, a scala astronomica questa rappresenta una distanza molto limitata e può interferire con numerosi fenomeni naturali. Tali oggetti/asteroidi possono dimostrarsi dannosi. Pertanto, è prudente conoscere ciò che ci circonda e identificare eventuali minacce tra di essi. Questo insieme di dati raccoglie l'elenco degli asteroidi certificati dalla NASA classificati come oggetti più vicini alla Terra.

Obiettivo dell'elaborato è boh

Scelte di design per la creazione del data set, eventuali ipotesi o assunzioni

Descrizione del dataset e relativa analisi esplorativa

Descrizione delle variabili

Variabile	Tipo	Contesto	Aspettativa	commenti
id	int	Valore di identificazione del singolo oggetto	Bassa	
name	Object	Valore di identificazione del singolo oggetto	Bassa	
est_diameter_min	Float	Misura minore "possibile" estimata del diametro	Media	
est_diameter_max	Float	Misura più alta "possibile" estimata del diametro	Media	
relative_velocity	Float	Misura della velocità dell'oggetto	Alta	
miss_distance	Float	Distanza tra la terra e l'oggetto	Media	
orbiting_body	Object	il corpo attorno cui orbita l'oggetto	Bassa	Il corpo è sempre la terra
sentry_object	Bool	ncluso in sentry - un sistema automatizzato di monitoraggio delle collisioni	Bassa	Sono tutti false
absolute_magnitude	Float	Misura della luminosità intrinseca	Media	
hazardous	Bool	True se l'oggetto è considerato pericoloso, false altrimenti	Alta	

Analisi esplorativa

L'analisi esplorativa del dato (exploratory data analysis, EDA) è di fondamentale importanza perché permette all'analista di conoscere a fondo il dataset sul quale lavora, stipulare o scartare ipotesi e creare dei modelli predittivi su basi solide. Si è deciso di suddividere questa fase nei seguenti punti: Comprensione del quadro generale, Preparazione, Comprensione delle variabili e Studio delle relazioni tra variabili

1. Comprensione del quadro generale

Il dataset ha dimensione di 90836 righe e 10 colonne, di seguito viene riportata una tabella descrittiva di alcuni tra i valori statistici fondamentali:

	id	est_diameter_min	est_diameter_max	relative_velocity	miss_distance	absolute_magnitude
count	9.083600e+04	90836.000000	90836.000000	90836.000000	9.083600e+04	90836.000000
mean	1.438288e+07	0.127432	0.284947	48066.918918	3.706655e+07	23.527103
std	2.087202e+07	0.298511	0.667491	25293.296961	2.235204e+07	2.894086
min	2.000433e+06	0.000609	0.001362	203.346433	6.745533e+03	9.230000
25%	3.448110e+06	0.019256	0.043057	28619.020645	1.721082e+07	21.340000
50%	3.748362e+06	0.048368	0.108153	44190.117890	3.784658e+07	23.700000
75%	3.884023e+06	0.143402	0.320656	62923.604633	5.654900e+07	25.700000
max	5.427591e+07	37.892650	84.730541	236990.128088	7.479865e+07	33.200000

Alucne osservazioni dei dati nella precedente tabella:

Dimensioni stimate

- La dimensione stimata degli oggetti celesti nel dataset varia notevolmente, con un diametro minimo che va da 0.000609 a 37.892650 e un diametro massimo che va da 0.001362 a 84.730541.

Velocità relativa

- La velocità relativa degli oggetti celesti è abbastanza varia, con una media di circa 48,066 unità di misura e una deviazione standard di circa 25,293. La velocità minima registrata è di circa 203.35, mentre la massima è di circa 236,990.13.

Distanza ravvicinata

- La distanza più vicina alla quale gli oggetti celesti si avvicinano è compresa tra circa 6,745.53 e 74,798,650 unità di misura, con una media di circa 37,066,550. Questo indica una vasta gamma di distanze ravvicinate.

Magnitudine assoluta

- La magnitudine assoluta degli oggetti celesti varia da 9.23 a 33.2, con una media di circa 23.53. Questo fornisce informazioni sulla luminosità intrinseca degli oggetti celesti nel dataset.

Variazione nei dati

- Le deviazioni standard relativamente alte in alcune colonne indicano una significativa variazione nei dati. Ad esempio, la deviazione standard elevata nella colonna "relative_velocity" suggerisce una grande variabilità nelle velocità relative degli oggetti celesti.

2. Preparazione

In questa fase si vuole iniziare a pulire il dataset in modo da continuare l'analisi. Alcune delle domande che aiuteranno a comprendere se il dataset contiene elementi da modificare sono:

1. **esistono variabili inutili o ridondanti?**

   Si, ad esempio è evidente che le variabili sentry_object e orbiting_body hanno lo stesso valore per ogni dato nel dataset, inoltre le variabili id e name servono entrmabe ad identificare un oggetto specifico. Per avere un dataset maggiormente ordinato e privi di variabili che non saranno oggetto di analisi, procederemo rimuovendo dal dataset òe varibili sentry_object, orbiting_body e name.

2. **Ci sono delle colonne duplicate?**

No, non vi è presenza di colonne duplicate.

3. La nomenclatura ha senso?

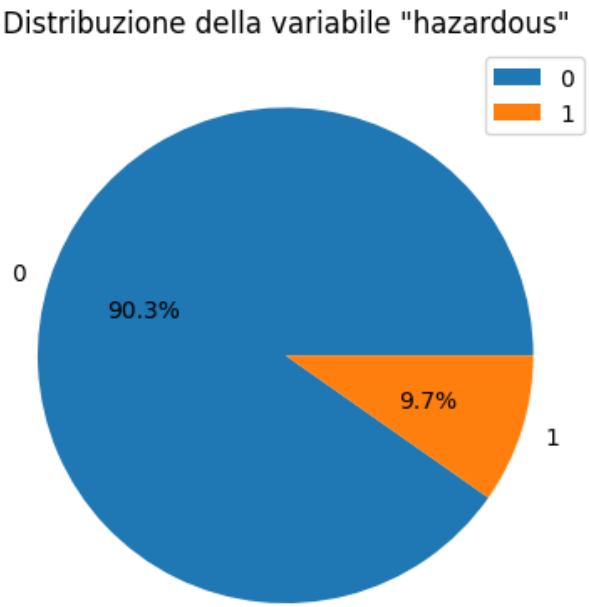
   Si, il modo in cui le variabili sono nominate rappresentano in modo sintetico il loro significato intrinseco.

4. Ci sono delle nuove variabili che vogliamo creare? Si, potrebbe essere utile studiare la media tra est_diameter_min e est_diameter_max, quindi procederemo aggiungendo questa variabile nel dataset.

3. Comprensione delle variabili (Analisi univariata)

Fase utile nel comprendere e descrivere le variabili di interesse.

hazardous



4. Studio delle relazioni tra variabili

Descrizione e motivazione dei modelli di machine learning scelti (almeno due modelli)

Esperimenti: esecuzione di almeno una modalità di validazione e stima delle misure di performance

Analisi dei risultati ottenuti

Conclusioni