Dataset Analyse & Model Building

Nasa Asteroids

H3 Hitema - Gerdy Jérôme

30 Octobre 2020

Objectifs : Les objectifs de ce projets sont d'analyser les données présentes dans le dataset **Nasa Asteroids**, et d'étudier la possibilité à l'aide de ces données et de recherche d'algorithme adapté, de prédire de manière la plus précise possible si un astéroïde est dangereux ou non pour la planète Terre.

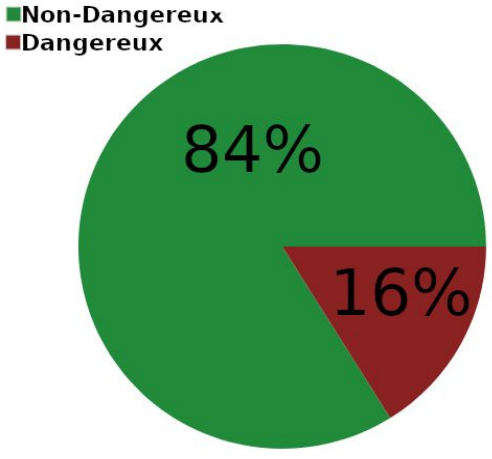
Partie 1: Présentation du dataset

Le dataset utilisé pour cette étude possède originellement 40 colonnes représentant différentes caractéristiques d'un astéroïde permettant de déterminer sa position, sa vitesse de déplacement, etc.

Column	Non-Null Count	Dtype
Neo Reference ID	4687 non-null	int64
Name	4687 non-null	int64
Absolute Magnitude	4687 non-null	float64
Est Dia in KM(min)	4687 non-null	float64
Est Dia in KM(max)	4687 non-null	float64
Est Dia in M(min)	4687 non-null	float64
Est Dia in M(max)	4687 non-null	float64
Est Dia in Miles(min)	4687 non-null	float64
Est Dia in Miles(max)	4687 non-null	float64
Est Dia in Feet(min)	4687 non-null	float64
Est Dia in Feet(max)	4687 non-null	float64
Close Approach Date	4687 non-null	object
Epoch Date Close Approach	4687 non-null	int64
Relative Velocity km per sec	4687 non-null	float64
Relative Velocity km per hr	4687 non-null	float64
Miles per hour	4687 non-null	float64
Miss Dist.(Astronomical)	4687 non-null	float64
Miss Dist.(lunar)	4687 non-null	float64
Miss Dist.(kilometers)	4687 non-null	float64
Miss Dist.(miles)	4687 non-null	float64
Orbiting Body	4687 non-null	object
Orbit ID	4687 non-null	int64
Orbit Determination Date	4687 non-null	object
Orbit Uncertainity	4687 non-null	int64
Minimum Orbit Intersection	4687 non-null	float64
Jupiter Tisserand Invariant	4687 non-null	float64
Epoch Osculation	4687 non-null	float64
Eccentricity	4687 non-null	float64
Semi Major Axis	4687 non-null	float64
Inclination	4687 non-null	float64
Asc Node Longitude	4687 non-null	float64
Orbital Period	4687 non-null	float64
Perihelion Distance	4687 non-null	float64
Perihelion Arg	4687 non-null	float64
Aphelion Dist	4687 non-null	float64
Perihelion Time	4687 non-null	float64
Mean Anomaly	4687 non-null	float64
Mean Motion	4687 non-null	float64
Equinox	4687 non-null	object
Hazardous	4687 non-null	bool

Certaines de ces colonnes sont cependant redondantes, notamment celles représentant la même valeur, avec simplement une unité différente (KM, Mètres, Miles, Feet, etc.). Seules les colonnes de l'unité Miles seront conservées de ce dataset.
Les colonnes **Orbiting Body** et **Equinox** ont également été retirées, car elles ne possédaient qu'une seule et unique même valeur chacune, leur présence n'était donc pas pertinente pour l'analyse, ainsi que pour la recherche d'algorithme.
Le dataset utilisé finalement pour cette étude possède donc 28 colonnes

En regardant la répartition des données, on voit que 84% des astéroïdes présents dans le dataset sont non-dangereux, et 16% le sont.



Il restera à déterminer si l'algorithme choisi au final trouvera des résultats similaires

Recherche du meilleur algorithme de prédiction

KNearestNeighbors

KNearestNeighors ou **K des plus proches voisins** est un des algorithmes les plus courants pour la classification

Après différents tests d'ajustement et d'optimisation de l'**accuracy** de l'algorithme, les 3 paramètres suivants sont ressortis:

- **n_neighbors = 5** : Nombre de voisins à utiliser par l'algorithme
- **weights = "distance"** : Fonction de poids à utiliser pour la prédiction. Le poids est définit par la distance entre les points (les "voisins"). Plus les voisins sont proches du point recherché, plus ils ont d'influence dans la prédiction
- **p = 1** : Correspond au paramètre de puissance pour la métrique de Minkowski
  Cette métrique a joué un rôle dans le domaine des théories de la relativité, qui est un des principes du modèle KNN

Le score de cet algorithme est d'environ **88.5%**, ce qui signifie que dans 88 cas sur 100, ce modèle arrive à prédire correctement la dangerosité d'un astéroïde.

SGD Classifier

Ce modèle est un classifieur linéaire, qui implémente l'apprentissage de la descente de gradient stochastique. C'est à dire que le modèle évalue pour chaque échantillon, la perte, et se réajuste à l'aide d'un taux d'apprentissage.

Les paramètres étant ressortis comme les plus optimisés pour ce modèle sont les suivants:

- **alpha = 5 :** L'alpha est utilisé pour le calcul du taux d'apprentissage et du terme de régularisation. Plus cette valeur est élevée, plus la régularisation est forte
- **n_jobs = 2** : Nombre de CPU à utiliser pour la réalisation du One Vs All. 2 Est le meilleur paramétrage dans notre cas
- **shuffle = False :** Ce paramètre précise si les données d'entraînement sont mélangées de nouveau après chaque boucle d'exécution. Dans notre cas, cela permet surtout de garder un résultat constant.

Son score est d'environ **91.4%** donc on est ici sur une prédiction juste dans 91 cas sur 100.

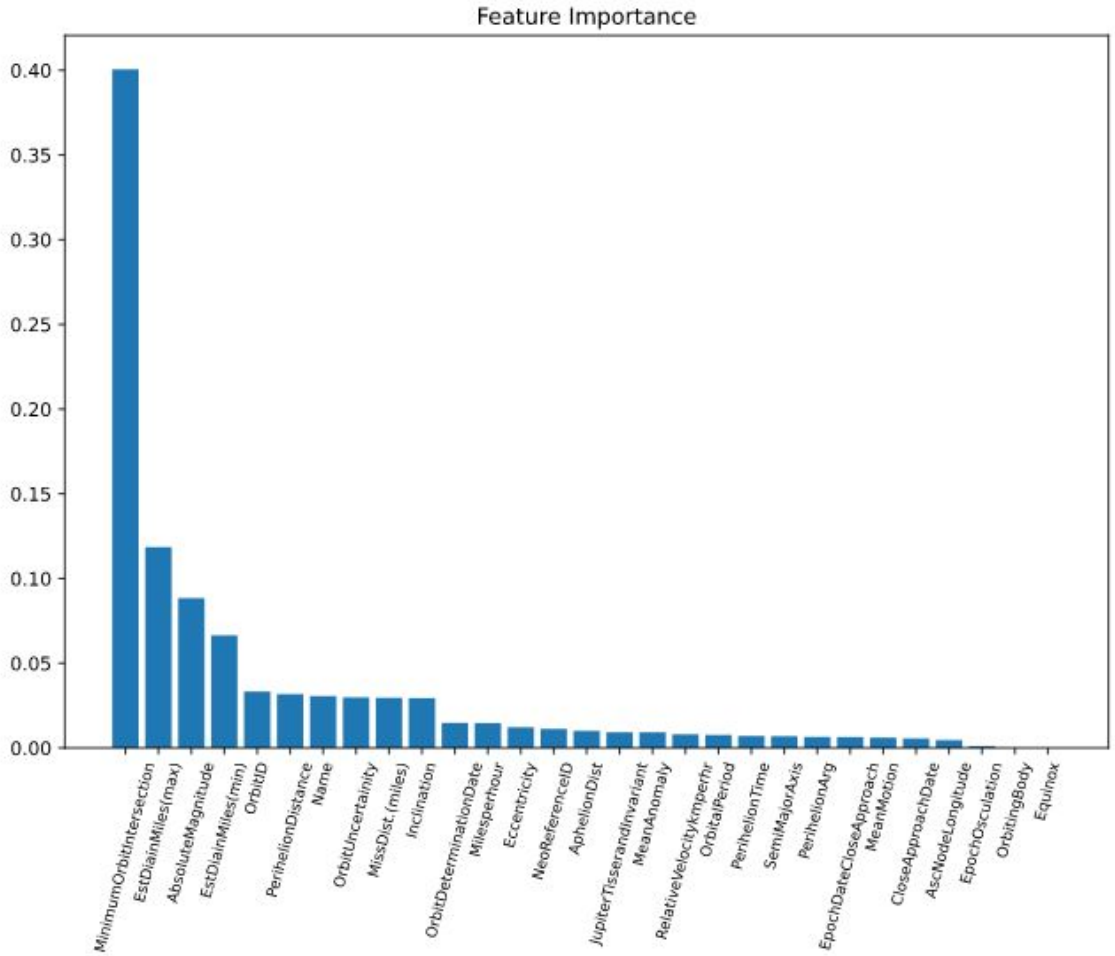
Random Forest Classifier

Random Forest, ou **forêt d'arbre de décision** est un modèle de classification, qui fait partie des modèles à apprentissage automatique. Sur chaque échantillon, on entraîne un arbre de décision, et chaque échantillon représenté par un arbre représentent en leur ensemble une forêt de décision.

Avec ce modèle, le score de prédiction est de **99.1%**, sans recherche d'hyperparamètre optimisé. Après de nombreux tests sur ces derniers, aucun n'améliore réellement le résultat. Donc dans 99 cas sur 100, l'astéroïde sera correctement déterminé dangereux ou non.

Model Réexecution

Après ces tests et analyses de ces 3 modèles, les 5 features (variables) impactant le plus le résultat on été trouvées, à l'aide de la **feature importance** :



- **MinimumOrbitIntersection**
- **EstDiainMiles(min)**
- **EstDiainMiles(max)**
- **AbsoluteMagnitude**

- **OrbitID**

Le principe du Model Reexecution est de réaliser la même expérience avec les 3 algorithmes, en ne lui donnant que ces variables pour réaliser leurs calculs.

Model reexecution de KNearestNeighbors

En réalisant une réexecution sur ce modèle en conservant unqiuement les 5 features citées au dessus, le résultat est désormais de **98.4%** de prédiction juste.

Model reexecution de SGD Classifier

Avec ce modèle, le résultat passe de 91.4% à **92.5%**. On constate donc que la différence avant/après le model reexecution est moins importante que pour KNN

Model Reexecution de Random Forest

Avec ce modèles, le résultat passe de 99.1 à **99.3%.**

On peut donc conclure sur la Model Reexecution, que seul le modèle KNN voit son accuracy augmenter très significativement.

Enfin, pour conclure sur cette étude, on remarque qu'une fois ré-entraînés avec les données les plus importantes, les trois modèles utilisés ici sont relativement très performants. Néanmoins, dans le domaine de l'aérospatial, les erreurs les plus infimes peuvent coûter très chères et des vies. Donc le modèle Random Forest est finalement le plus approprié pour la prédiction de la dangerosité d'un astéroïde, même si une double vérification humaine est sûrement nécessaire pour ne pas risquer le pire.