Pratiques d'exploration de données pour quantifier les risques d'impacts d'astéroïdes - Une approche prédictive -

Rapport Methode de Simulation Numerique

Dima Elisabeta Iulia, Rakotozafy Njariniaina Emmanoella

Licence 3 Informatique, Université Côte d'azur Campus Valrose

May, 2022

Résumé

Ce projet analysera certaines données dont nous disposons sur les astéroïdes en examinant des critères comme la probabilité d'une collision avec la Terre et la précision de la prédiction d'un impact ou classification d'un astéroïde.

Figure 1: Dangereux astéroides. Sommes-nous prêts à éviter une catastrophe ?

1. Introduction

La surface de la terre est une cible parfaite des météorites. Mais selon les statistiques, il ne tombe d'astéroïdes de quelques dizaines de mètres de diamètre que chaque siècle.
Des lourds impacts sont constatés après le passage d'un astéroïde, c'est pour cette raison que notre projet vise à sensibiliser en fournissant des algorithmes pertinents et des résultats précis concernant d'éventuels astéroïdes impactant, qui pourraient constituer une menace à une certaine échelle (c'est-à-dire une ville, un pays, un continent ou même notre planète entière).
Pour bien mener notre étude, nous utilisons une base de données de NASA JPL auquel notre travail se déroule en trois parties : l'analyse des données, l'apprentissage supervisé et l'apprentissage non–supervisé avec des techniques de validation.

1.1. Problématique

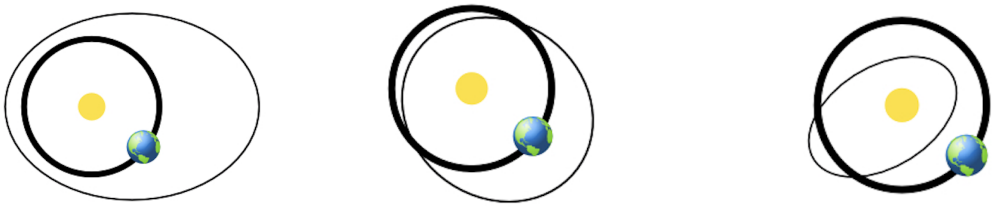
Comment peut-on prédire qu'un astéroïde présente un risque d'impact sur la terre ? A partir de quelle echelle un astéroïde est-il consideré comme dangereux ?

2. Théorie

Pour mieux comprendre la base de données et la corrélations entre les colonnes, on va spécifier les plus importantes caractéristiques des astéroides:

- H - magnitude absolue (luminosité intrinsèque d'un objet vue à distance de 1 unité astronomique de la Terre. 1 UA $= 1.495 * 10^8 km$)
- a - albedo, le pouvoir réfléchissant d'une surface (0-absorption totale, 1-réflection totale).
- diamètre (km)
- q - perihelion (distance minimale du Soleil)
- Q - aphelion (distance maximale du Soleil)
- NEO - Near Earth Asteroid, q<1.3 UA
- MOID - Earth Minimum Orbit Intersection Distance

En analysant ces notions, on peut déduire assez facilement les astéroïdes qui peuvent provoquer des inquiétudes. En effet, le JPL (*Jet Propulsion Laboratory*) de NASA spécifie les paramètres d'un asteroide PHA (Potentiellement Hasardeux) sont $MOID \leq 0.05$ UA (7 480 000 km) et $H \leq 22$.



(a) Classe AMO(Amor)
a > 1.0 AU; 1.017 AU < q < 1.3 AU

(b) Classe APO(Apollo)
a > 1.0 AU; q < 1.017 AU

(c) Classe ATE(Aten)
a < 1.0 AU; Q > 0.983 AU

Figure 2: Classes des asteroides avec un risque d'impact elevé.

Chaque astéroide s'inscrit dans une certaine orbite définie par une ellipse avec un degrée d'elongation.

$$e = \frac{\sqrt{a^2 - b^2}}{a^2} \quad (1)$$

e: excentricité de l'ellipse
a: semi-axe majeur
b: semi-axe mineur

$$q = a * (1 - e)$$
$$Q = a * (1 + e)$$

La **Figure 2.** montre les ellipses ayant des coordonnées qui peuvent intersecter la Terre. En plus, dès que la gravitation d'un astéroide est incomparable avec celle de la Terre, son MOID devient un facteur parfois décisive pour qu'il soit attiré par la Terre et entre dans son champ gravitationnel. Les asteroides qui ont PHA=Y sont généralement de type APO, AMO ou ATE (ces affirmations seront aussi prouvées par l'analyse de la base de données).

Il est important aussi de prendre en compte la magnitude absolue et l'albedo.On peut mesurer le diamètre d'un astéroide du façon indirect par la formule:

$$d = 10^{3.1236 - 0.5 * \log_{10}(a) - 0.2 * H} \quad (2)$$

Concernant le diamètre d'un astéroïde, l'ESA (*l' Agence Spatiale Européenne*) affirme que "Les astéroïdes de large diamètre ne présente aucun risque, aussi que les astéroïdes de taille très petites(<10m). Le principal défi provient de la population d'objets de taille moyenne, allant de dizaines à centaines de mètres de diamètre." Cette affirmation sera l'une des raisons pour laquelle on va réduire notre base de données.

3. Travaux connexes

- A probabilistic asteroid impact risk model: assessment of sub-300 m impacts: modèle PAIR, basé sur le taux de fréquence d'impact publiés avec des outils d'évaluation des conséquences de pointe, appliqués dans un cadre de Monte Carlo (astéroide ¡= 300m diamètre) (lien)
- Quantifying the Risk Posed by Potential Earth Impacts:comparaison et catégorisation des nombreuses solutions d'impact potentiel découvertes, basée sur une nouvelle échelle de danger qui décrira le risque posé par un impact potentiel (lien)
- Deflection driven evolution of asteroid impact risk under large uncertainties: basée sur une approche statistique(échantillonnage des distributions de propriétés orbitales et physiques, déviation d'impacteur cinétique préconçues sont prises en compte) (lien)

Le point en commun de ces travaux avec le notre est qu'ils recherchent tous à étudier les risques d'impact d'astéroides. Mais notre travail se diffèrent des leurs en nous basant sur l'étude des astéroides potentiellement dangereux (PHA qui est notre Y prédiction).

4. Jeu de données et fonctionnalité

Notre base de données de taille initiale 958524 valeurs * 45 colonnes a été réduite à 11524 valeurs * 37 colonnes. On a eliminé les astéroides avec des diamètres >2 km ainsi que les données de n'importe quelle colonne avec des valeurs nulles.
Nous tenons à péciser aussi que nous n'avons pas normalisé notre base de données mais on a binarisé les classes, ce qui fait que le nombre de colonnes a été augmenté 8 de plus. Nous avons constanté la présence de données aberrantes (Ceres - 952 km diametre). Les éléments qui nous importe le plus sont: le diamètre, le pha, le néo, le H, le moid, le perihelion, l'albedo et les classes.

Notre base de données contient que des astéroides qui se trouvent dans notre Système Solaire. La plus part d'entre eux proviennent de la Ceinture principale (MBA- Main Belt Asteroids).

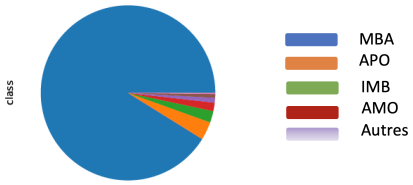


Figure 3: Diagrame circulaire des classes.

On a analysé les corrélations entre les bases de données (qui peuvent nous aider aussi à avoir des bonnes précisions des modèles). La **Figure 4.** montre une matrice de chaleur (ici juste pour les plus importantes caractéristiques). Il y a des corrélations positives entre le PHA, le H et le NEO et une corrélation négative entre MOID et NEO (un MOID élevé donne un NEO nul).

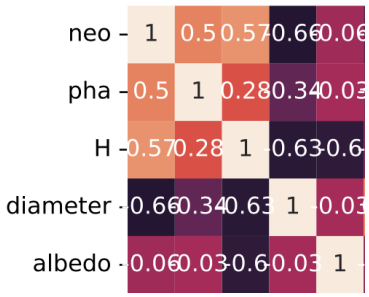


Figure 4: Correlations des données

5. Méthodes

Pour les classifications des données, les modèles utilisées sont: Kernel Trick, SVM, Reapeted Random sampling, Random forest Classifier, K-fold Cross-validation et K-stratified fold.

- Kernel Trick: Avec cete méthode, les classes peuvent devenir linéairement séparables dans un espace de caractéristiques de dimension supérieure.
- SVM: L'objectif est de trouver un hyperplan qui vise à séparer, si possible, les deux classes.
- Repeated random subsampling: divise l'ensemble de données de manière aléatoire en échantillons d'apprentissage et test de façon repetée (nombre d'itérations).
- Random forest: (ou forêt aléatoire) C'est un algorithme sophistiqué, dans le domaine du machine learning. Il permet d'obtenir une prédiction fiable, grâce à son système de forêt d'arbres décisionnels.
- K-fold Cross-validation: Elle résulte généralement sur un modèle moins biaisé. Pour cause, elle permet d'assurer que toutes les observations de l'ensemble de données original aient la chance d'apparaître dans l'ensemble d'entraînement et dans l'ensemble de test.
- StratifiedKfold: (ou validation croisée stratifiée) C'est une extension de la technique de validation croisée utilisée pour les problèmes de classification. Il maintient le même rapport de classe à travers les plis K que le rapport dans l'ensemble de données d'origine.

6. Expériences et résultats

6.1. Séparation des données

1. Kernel Trick: A l'aide de cet algorithme supervisé, on a essayé de faire une séparation linéaire avec une dimension et deux dimensions pour observer si la distance minimale de soleil (potentiellement la distance minimale de la terre) appelée *perihelion* permet d'identifier qu'un astéroide pourrait devenir un *PHA*. Selon le résultat, il est impossible d'avoir une séparation linéaire puisque les données sont tellement randomisées de telle sorte qu'on ne peut pas savoir de quelle distance entre un point de l'ellipse et la terre, un astéroide peut être classé NEO ou PHA.

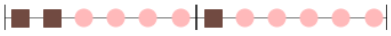


Figure 5: Points de données d'origine

2. SVM: Dans cette partie, on a etudié tout d'abord la corrélation entre le diamètre et le MOID (distance minimale de l'intérsection avec la terre). On remarque un groupement d'astéroides de taille de diamètre très grande et dont leur MOID aussi est assez grand.
   Par conséquent, les astéroides qui sont près de la terre et qui sont probablement dangereux, classés en PHA n'ont pas de grande taille, parce que la probabilité d'un astéroide de taille grande et un MOID faible est minime.
   À l'aide d'algorithme suppervisé SVM, on a essayé de séparer les données en utilisant un hyperplane et un paramètre C=10 de contrôle de violation de la marge.

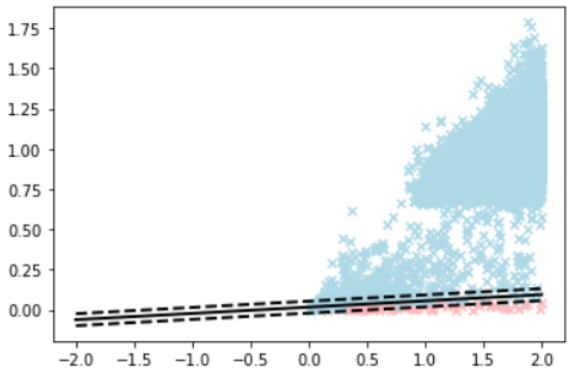


Figure 6: Base de données avec un hyperplane séparateur (SVM)

6.2. Génération des données d'entrainement, de test et de validation - Calcul de la précision moyenne et la variance

Dans cette partie, nous allons diviser notre base de données en trois parties pour génerer des données d'entrainement, de test et de validation en utilisant des différents algorithmes afin d'avoir des résultats de précision. Pour cela, deux options sont possibles: soient on mélange les données, soit on les laisse à leur forme originale.

1. SVM- Random subsampling avec d=2:
   On a pris un nombre d'échantillons de 100 (distrubué de façon randomisé) et on a utilisé l'algorthme SVC (avec un kernel linéaire). Le résultat nous montre des valeurs assez précises, même très précises parfois qui est justifié par le fait qu'on a beaucoup de données et on observe que plus on varie les valeurs de C, plus on obtient des résultats exagerés (On obtient les meilleurs resultats pour C_i=1000).

```
Accuracy training   [ 94.  94. 100.]
Accuracy testing    [92. 92. 98.]
Accuracy validation [98. 98. 98.]
The accuracy in the studied dataset is 98.00
La précision moyenne est  98.00
La variance de précision est 0.00
```

2. SVM-Kernel avec d=2:
   Il s'agit de diviser les données: soit de façon randomisé, soit de façon originelle et on utilise les trois types de kernel: Linéaire, rbf, polynomiale. Cette méthode aboutit également à un résultat précis tout comme avec le Repeated Random Subsampling.

```
The accuracy in the studied dataset is 99.48
La précision moyenne est  99.48
La variance de précision est 0.00
```

3. Kfold Cross-validation avec d=10:
   Dans la première instance, nous n'avons pas utilisé que les données originelles: 50% entrainement, 25% test et 25% validation. Et dans la deuxième instance, on a mélangé les données dont les données echatillonnées sont repartis comme suit: 40% entrainement, 30% test et %30 validation. Après avoir fait 1 itération en utilisant 3 hyperparamètres de contrôle, on a réalisé pour 10 itérations avec 5 valeurs de C différentes. Cette méthode peut nous permettre d'avoir un bon résultat dans certaines itérations (moyenne de précison elèvée et variance trop faible).

```
La précision sauf randomiser est 98.02
La précision (random) est 96.67
La précision moyenne est  97.34
La variance de précision est 0.46
```

4. StratifiedKFold:
   Toujours avec 10 itérations comme avec le Cross-validation, cette méthode nous permet d'avoir des bons résultats seulement pour certaines itérations (les variances ont même des valeurs entre 0.02 et 16.09).

```
La précision sauf randomiser est 98.02
La précision (random) est 98.33
La précision moyenne est  98.18
La variance de précision est 0.02
Iteration  6
```

6.3. Précisions des modèles: SVC, KNeighbours, RandomForestClassifier

(a) Score des modèles:
Pour cette dernière étape de validation et d'apprentissage, on va utiliser toutes les colonnes de notre base de données en utilisant le SVC, KNeighbours, RandomForestClassifier pour 10 itérations. Pour cela, on sépare les données on utilise 30% de données de test en les distribuant de façon randomisés.
Mais dans le but d'avoir un meilleur résultat, on va proceder à un pré-traitement de données avant de faire l'algorithme.Ainsi, le Stantardscaler est utilisé pour normaliser les caractéristiques tout en adaptant le modèle. A la sortie, on obtient des point variés entre 0 et 1 après la transormation. L'initialisation des modèles est nécéssaire pour avoir une prédiction et un calcul de score très productibles.

	Score
KNeighborsClassifier	0.986119
SVC	0.988143
RandomForestClassifier	0.998554

Figure 7: Score des modèles

(b) Matrice de confusions des modèles:
Après avoir analysé chaque matrice de confusion, on observe le fait que le RandomForestClassifier donne une meilleur prédiction par rapport aux autres modèles.

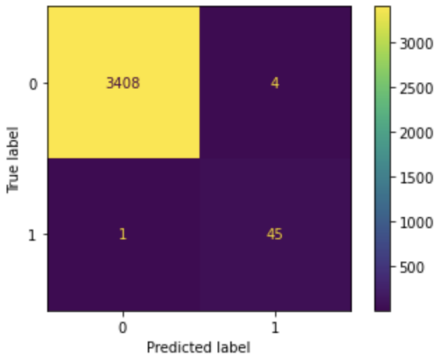


Figure 8: Matrice de confusion (RFC).

(c) Précision, moyenne et variance des modèles:
Dans cette parties, nous allons calculer la précision, la moyenne et la variance de chaque modèle pour 10 itérations pour savoir lequel est le meilleur algorithme d'entre les trois.Et compte tenu des résultats de calcul, c'est l'algorithme RandomForestClassifier qui est le meilleur car il a un taux de précision grand et une faible variance par rapport aux autres algorithmes.

```
La precision de RFC pour 10 iterations:
[1.         1.         1.         1.         1.         1.
 1.         1.         0.98784722 0.96788194]
La moyenne (RFC): 0.9955729166666666
La variance (RFC): 9.832688319830261e-05
```

Remarque: On utilise un model de type *Baseline* pour contextualiser les résultats des trois modèles entraînés. Notre modèles intègrent la randomisation, donc il est important de définir une valeur de départ afin que nos résultats soient reproductibles.

7. Discussion

Après avoir étudié notre base de données en utilisant tous les classificateurs possibles pour savoir lequel donne une bonne prédiction, prouvée par le fait d'avoir une grande précision et une variance très faible, on a constaté que c'est RandomForestClassifier qui est le meilleur dans notre cas. Cette étude nous a permis aussi de savoir les caractéristiques importants de notre base de données:le Moid, le H-magnitude absolue, la classe APO, le diamètre, qui nous ont aidé à bien mener ce travail. Nos résultats pratiques sont confirmés par la partie théorique de notre projet (selon l'ESA: $MOID \leq 0.05$ UA et $H \leq 22$. $\implies$ $PHA = 1$)

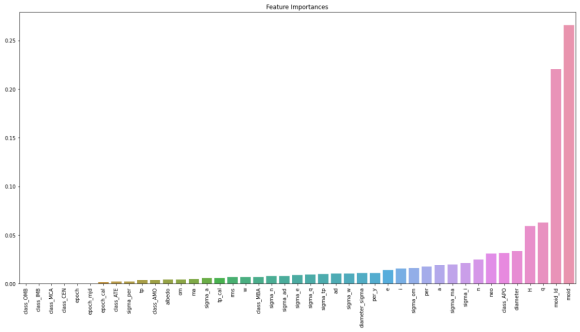


Figure 9: Histogramme des importances des caractéristiques de données

8. Conclusion

Ce projet nous a vraiment permis d'explorer de façon approfondi les élements et les caractéristiques à prendre en compte pour prédire quels types d'astéroïdes sont hasardeux. Puisque certains astéroïdes peuvent entrer en collision avec la Terre, ils sont également importants pour avoir considérablement modifié la biosphère terrestre dans le passé.Ils continueront de le faire à l'avenir...

References

1. https://www.kaggle.com/datasets/shrutimehta/nasa-asteroids-classification
2. https://www.esa.int/ESA_Multimedia/Images/2019/04/Detection_What_s_the_risk
3. https://www.businessinsider.com/asteroid-sizes-that-can-damage-cities-states-planet-2018- 6?op=1r=USIR=Tan-asteroid-more-than-half-a-mile-wide-would-start-to-have-global-implications-11
4. https://www.tameteo.com/actualites/actualite/dangereux-asteroides-sommes-nous-prets-a-eviter-une-catastrophe-terre.html
5. https://www.researchgate.net/publication/322310362_Automatic_Design_of _Missions_to _Small_Bodies
6. https://cneos.jpl.nasa.gov/about/neo_groups.html
7. https://www.esa.int/ESA_Multimedia/Images/2018/06/Asteroid_danger_explained
8. https://cneos.jpl.nasa.gov/glossary/h.html
9. https://cneos.jpl.nasa.gov/glossary/albedo.html
10. https://cneos.jpl.nasa.gov/tools/ast_size_est.html
11. https://cneos.jpl.nasa.gov/about/neo_groups.html
12. https://thecuriousastronomer.wordpress.com/2014/06/26/what-does-a-1-sigma-3-sigma-or-5-sigma-detection-mean/

9. Contributions

- Code: Iulia et Emmanoella
- Analyse de données: Iulia et Emmanoella
- Rapport sur Latex: Iulia et Emmanoella
- Présentation/ poster: Iulia et Emmanoella