Projet de Fouille De Donnée

Prédiction des goûts cinématographiques et musicaux

Liang GUAN
Binta DIABIRA

Introduction

Le sujet que l'on a choisi pour ce projet, est la prédiction des goûts cinématographiques d’une personne en fonction de ses goûts musicaux.Avant de choisir le sujet on s’est d’abord mis à la recherche de données intéressantes pour ensuite choisir le type de sujet(clustering, prediction, classement.....) en fonction des données.On a trouvé sur kaggle.com les résultats d’un test de personnalité mbti effectué sur 8600 personnes ainsi que les 50 derniers postes qu’elles ont publiés sur le internet. On a pensé à un sujet : prédire les résultats du test à partir du contenu de leurs poste. On s’est mis d’accord pour choisir une étude sociale, et a continué nos recherches pour voir s'il n’y avait pas une base de données encore plus intéressante. Puis on est tombé sur une étude sur les jeunes slovaques, cette étude contient plus d’informations que celle trouvée précédemment on s’est dites que l’on pourrait faire plus de chose avec ses données, si l’on avait le temps.

Base de donnée

La base de donnée est le résultat d’un sondage effectué sur 1010 jeunes Slovaques âgé de 15 à 30 ans. Ils ont été interrogé sur les sujets suivants :

- Préférences musicales
- Préférences cinématographiques
- Hobbies et passe-temps
- Peurs
- Hygiène de vie
- Personnalité et opinions
- Habitudes de consommation Données démographiques

Les réponses aux questions sont soit des entiers, soit des chaînes de caractères. On a demandé aux personnes interrogées de donner une note allant de 1 à 5, à chacun des types musicaux suivants:

- Slow songs or fast songs
- Dance
- Folk

- Country
- Classical music
- Musical
- Pop
- Rock
- Metal or Hardrock
- punk
- Hiphop,Rap
- Reggae Ska
- Swing Jazz
- Rock n roll
- Alternative
- Latino
- Techno, Trance
- Opera

Elles ont également noté les films suivants(de 1 à 5) :

- Horror
- Thriller
- Comedy
- Romantic
- Sci-fi
- War
- Fantasy/Fairy tales
- Animated
- Documentary
- Western
- Action

La base donnée étant un fichier CSV de 1011 lignes nous n’avons pas jugé utile d’utiliser un SGBD. On essaye donc de prédire la note donnée à un type de film en se basant sur les notes données aux différents types musicaux.

Difficultés rencontrés

Au départ on ne comptait pas faire de prédiction, on voulait faire du clustering en regroupant les gens selon leurs musiques préférées, puis vérifier si les personnes avec les mêmes goûts musicaux présentaient d'autres similarités(goûts cinématographiques, position géographique, habitudes alimentaires…), on a pensé à utiliser k-means pour regrouper les gens selon leurs types de musiques, puis faire un autre clustering sur les goûts cinématographiques(dans un premier temps puis avec les autres données) et ensuite comparer les clusters.Nous doutions de cela soit la bonne approche, le clustering soulevait plusieurs questions : fallait-il faire autant de clusters que de type de musique(18) ou regrouper certains types musicaux par exemple(par exemple rock et hardrock…) ? qu'elle distance utiliser ?Après discussion avec les enseignants on a convenu qu'il valait mieux utiliser l'apprentissage supervisé et prédire les notes des personnes interrogées.

Réalisation

On parse d'abord le fichier CSV pour obtenir les notes données aux différentes catégories musicales et à un type de film, il arrive que les personnes interrogées n'aient pas attribué de note à un certain type de musique dans ce cas on insère un trois, la note moyenne, à la place de les données manquantes.
On se retrouve avec un X de la forme :

rock	hiphop	opera	…
1	5	3	…
5	4	3	...

….

et un Y de la forme :
film d'horreur(ou autre)

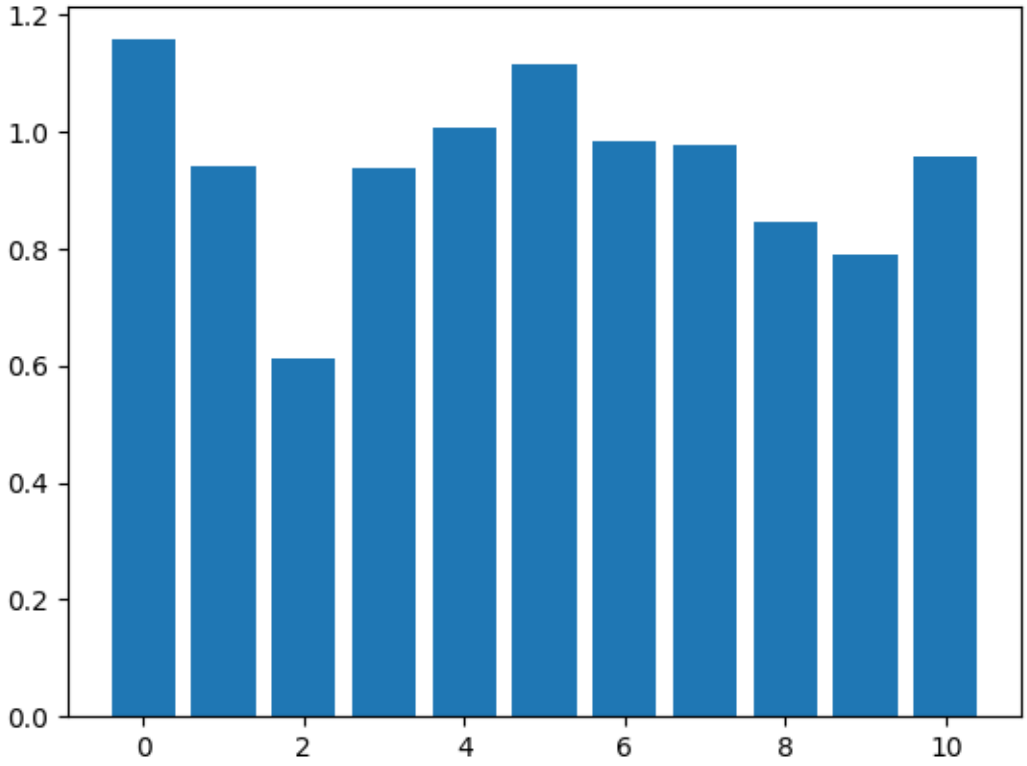
5
6
…

Chaque ligne de X et Y correspond aux notes d'une personne.On a ensuite

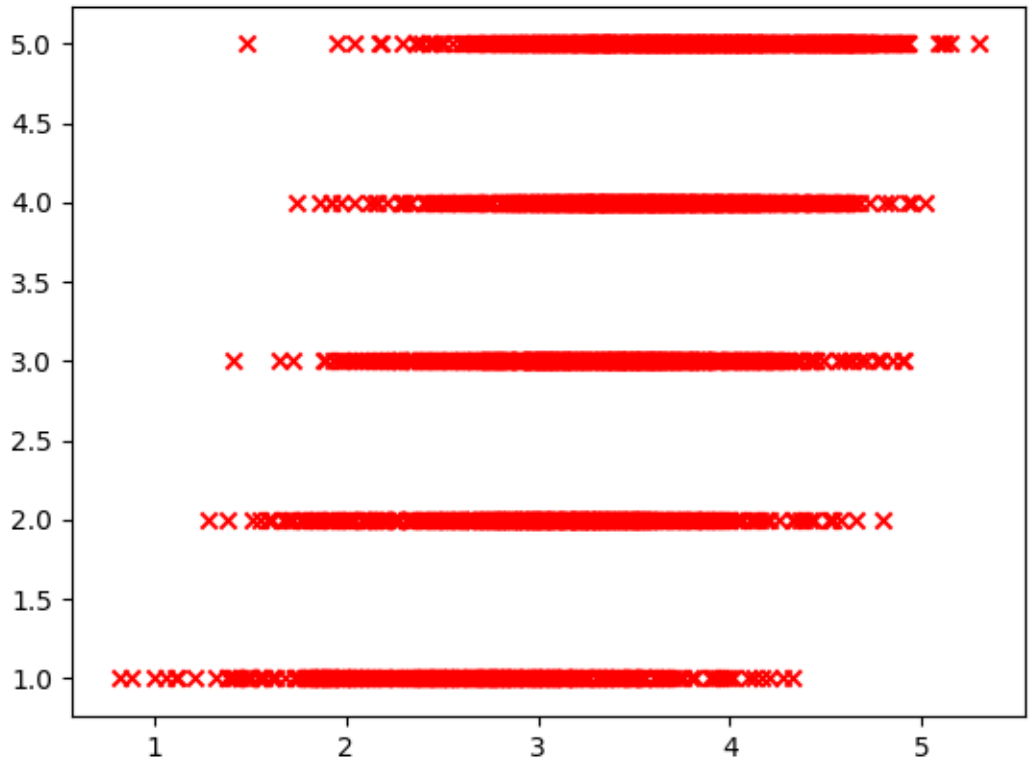
utilisé l'implémentation de la régression linéaire de la librairie sklearn de python, on a divisé les données en un ensemble d'entraînement, et un ensemble de tests sur lequel effectuer les prédictions.On utilise ensuite une métrique pour déterminer l'efficacité du modèle(MAE).

Résultat

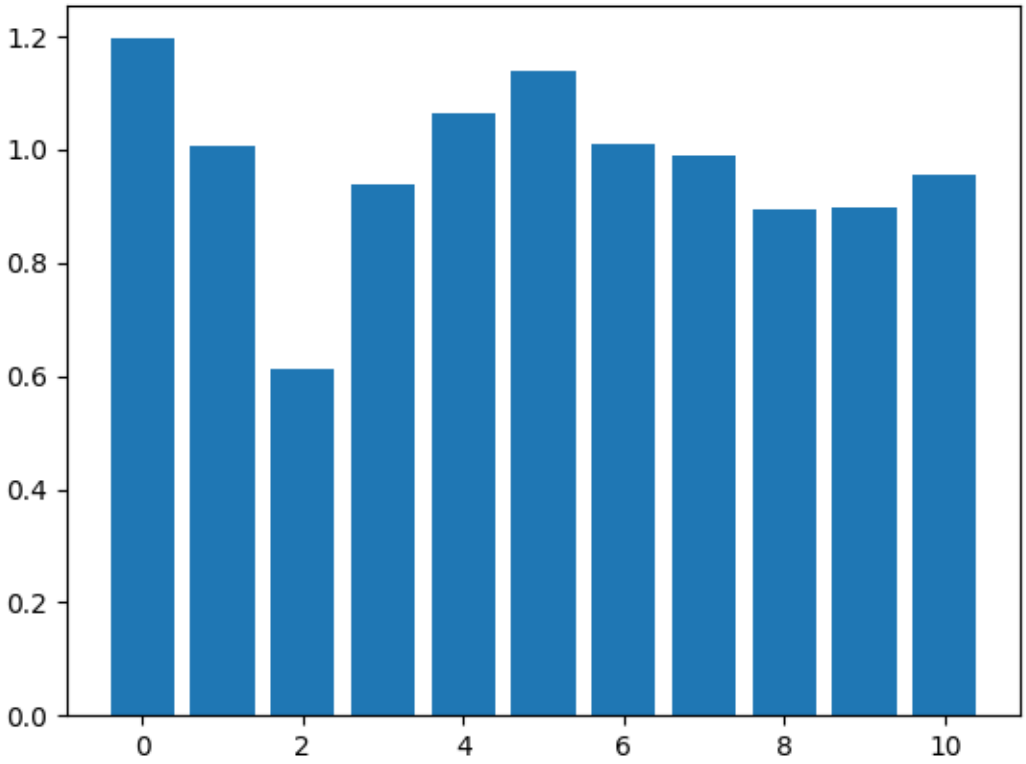
Valeurs des MAE pour les prédictions des notes des dix type de films par régression linéaire :



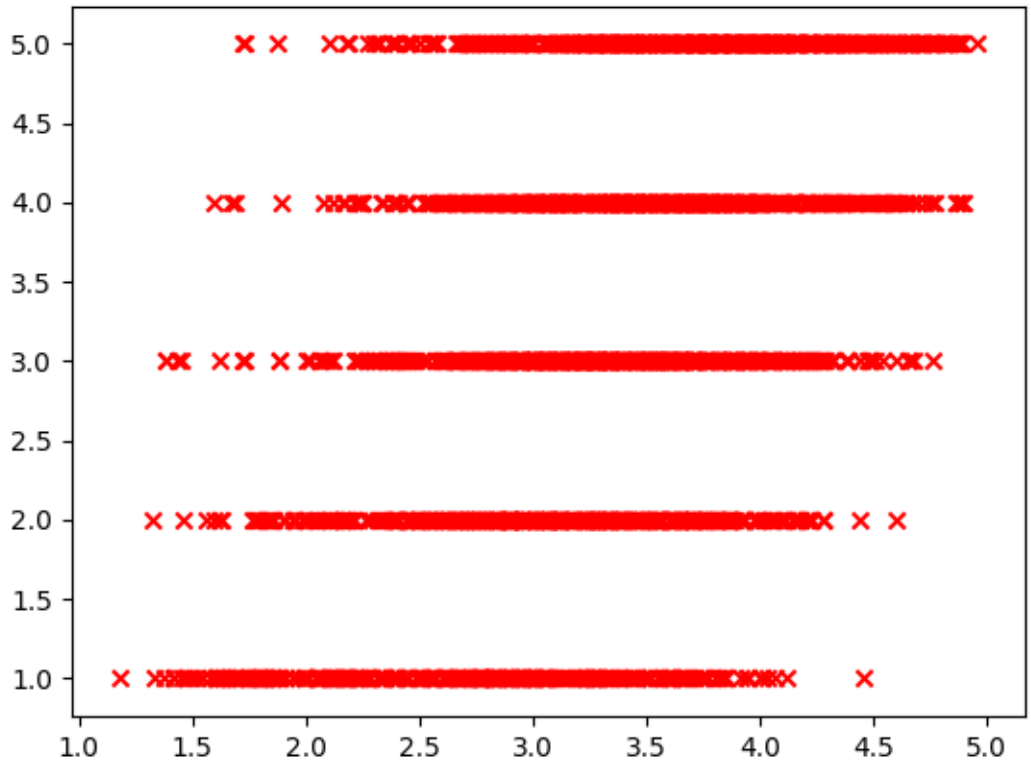
Moyenne : 0.93919859716



Valeurs des MAE pour la prédiction des notes des dix catégories de film avec les random forest :



Moyenne : 0.973295561487



Que ce soit en utilisant la régression linéaire ou les random forest, le mae est en moyenne de 0.9 la différence entre les prédictions et les notes réelles sont donc inférieurs à 1.

Conclusion

On a un modèle assez fiable, capable de prédire la note d’une personne pour une catégorie de film en se basant uniquement sur ses goûts musicaux avec une marge d'erreur inférieure à 1. La régression linéaire offre un modèle légèrement plus précis que les forêts aléatoires.

Répartition du travail

Projet fait en binôme, lors de rencontres à la fac.En général sur machine séparer et quelquefois sur une machine. Lorsqu’un problème est rencontré l’un de nous essaie de le résoudre pendant que l’autre avance ou s'il n’est possible d’avancer sans résoudre le problème en question, chacun essaye de le résoudre de son côté, puis la meilleure solution trouvée est sélectionné.