# Paramètres importants pour déterminer la même chanson

Nombre de jours depuis le dernier passage en même chanson

Nombre d’émissions depuis le dernier passage en même chanson (trop galère)

Nombre de jours depuis le dernier passage pour des points

Nombre d’émissions depuis le dernier passage pour des points (trop galère)

Cluster

Nbpassages en même chanson

Avec

* calculée à partir de l’histogramme des valeurs moyennes des écarts PMC – DP.
* calculée à partir de l’histogramme des dernières valeurs des écarts PMC – DP
* calculée à partir de l’histogramme des valeurs moyennes des écarts PMC – DPMC
* calculée à partir de l’histogramme des dernières valeurs des écarts PMC – DPMC
* donnant le nombre de passages de la chanson

Actuellement :

## Optimisation de f

On détermine le nombre de bins de l’histogramme en faisant tourner le programme optimisationEchf.py sur les données de 01-01-2021 à 31-08-2021.

Résultats :

|  |  |  |
| --- | --- | --- |
| Bins | Moyenne | Médiane |
| 10 | 154.3 | 144 |
| 11 | 139.4 | 92 |
| 12 | 134.5 | 87 |
| 13 | 130.0 | 83 |
| 14 | **130.8** | **82** |
| 15 | 135.1 | 92.5 |
| 16 | 135.5 | 96.5 |
|  |  |  |
|  |  |  |

## Optimisation de g

On détermine le nombre de bins de l’histogramme en faisant tourner le programme optimisationEchg.py sur les données de 01-01-2021 à 31-08-2021.

Résultats :

|  |  |  |
| --- | --- | --- |
| Bins | Moyenne | Médiane |
| 10 | 151.8 | 121.5 |
| 13 |  |  |
| 14 |  |  |
| 15 |  |  |
| 16 |  |  |
| 20 |  |  |

## Optimisation de h

On détermine le nombre de bins de l’histogramme en faisant tourner le programme optimisationEchh.py sur les données de 01-01-2021 à 31-08-2021.

Résultats :

|  |  |  |
| --- | --- | --- |
| Bins | Moyenne | Médiane |
| 10 | 188.3 | 171 |
| 11 | 186.4 | 174.5 |
| 12 | 183.2 | 168.5 |
| 13 | 182.4 | 175.5 |
| 14 | 180 | 161.5 |
| 15 | 177.9 | 150.5 |
| 16 | **173.7** | **139** |
| 17 | 175 | 146 |
| 18 | 173.9 | 146.5 |
| 19 | 175.2 | 151 |
| 20 | 170.7 | 148 |
| 21 |  |  |

## Construction des clusters

On remarque que les chansons qui tombent en même chansons ne tombent pas à la même fréquence. On cherche donc à intégrer ce phénomène au modèle en construisant des clusters de chansons qui ont le même comportement entre elles. On définit donc les variables qui comptent le nombre de passage dans chaque catégorie {50, 40, 30, 20 ou 10, MC, 20k}.

### Etude des échantillons

#### Pour n=3 clusters

|  |  |  |
| --- | --- | --- |
| **20k (595 chansons)** | **20k parfois (1104 chansons)** | **Mêmes chansons (506 chansons)** |
| Pourquoi c’est beau | Mme Pavoshko | J’ai dix ans |
| Ninja | Bonnie and Clyde | Capri c’est fini |
| Nouveau western | Le jour se lève | Tombé sous le charme |
| Quand tu m’aimes | Le travail c’est la santé | Si j’étais un homme |
| Respire | La neige au Sahara | Pas là |
| Mourir d’aimer | Puisque tu pars |  |
| Comme ils disent | Femme d’aujourd’hui |  |
| Au nom de la rose | Faut qu’j’travaille |  |

#### Pour n=4 clusters

|  |  |  |  |
| --- | --- | --- | --- |
| **Peu fréquentes (923 chansons)** | **MC fréquentes (329 chansons)** | **20k (528 chansons)** | **MC peu fréquentes (425 chansons)** |
| Polka | **La vie par procuration** | **Je viens du sud** | Travel the world |
| Aimer à mort | Le chant des sirènes | Avec classe | Ca ira mon amour |
| Partir là-bas | **Divine idylle** | Le mal aimé | **Les bêtises** |
| **Elle a fait un bébé toute seule** | **En rouge et noir** | Je reviendrai à Montréal | Les poèmes de Michelle |
| Musicienne | **L’Aziza** | Allez reste | Les filles de mon pays |
| L’équipe à Jojo | **Je suis un homme (Zazie)** | Miami Vice | Mamy Blue |
| Le casse de Brice | **Capri c’est fini** | La tendresse (Guichard) | Les princes des villes |
| Je t’aime à la folie | Derrière l’amour | Ton combat | Pour la vie |

#### Pour n=5 clusters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **0 : ? (912)** | **3 : 20k parfois (346)** | **2 : MC peu freq (224)** | **1 : MC freq (207)** | **4 : 20k souvent (516)** |
| Jolie nana | *Tous les bateaux tous les oiseaux* | **Elle habite ici** | **Cendrillon** | *La baie* |
| *La belle de Cadix* | *Dis-lui toi que je t’aime* | ***Sans contrefaçon*** | **Ne partez pas sans moi** | *Emma* |
| *Est-ce que c’est ça* | *La neige au sahara* | **On écrit sur les murs** | **Seul** | *Sous les jupes des filles* |
| Ma benz | *Pas besoin de toi* | **Tatoue-moi** | **Non non non (Ecouter Barbara)** | *Parlez-moi de lui* |
| Syracuse | *Coco Câline* | Alors on danse | **Le blues du businessman** | *La vie du bon côté* |
| Les derniers seront les premiers | Bubble Star | **Lili voulait aller danser** | **Belle demoiselle** | *Alors alors* |
| Amies ennemies | **Ta main (Grégoire)** | Elle s’ennuie | **J’te l’dis quand même** | *Chez moi* |
| N’y pense plus | Nue | *Colore* | **Entre nous** | *C’était bien (au p’tit bal perdu)* |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **41 MC** | **46 MC** | **93 MC** | **207 MC** | **23 MC** |
| *203 20k* | *171 20k* | *70 20k* | *0 20k* | *493 20k* |
| 668 random | 129 random | 61 random | 0 random | 0 random |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *21,7%* | *18,2%* | *7,8%* | *0%* | *52,6%* |

# Idée : classificateur catboost ou xgboost ou NN

Idée : on prend un dataset de lignes aléatoires avec comme features :

* Titre
* Artiste
* Année
* Décennie
* clusterID
* DeltaDate
* DeltaDateMC
* DeltaDateMCMA (même artiste : si Là-bas passe, pas d’Encore un matin le lendemain)
* Taux de choisi

Labels :

1. Passe pas
2. MC
3. 50
4. 40
5. 30
6. 20
7. 10
8. 20k

Génération du dataset : Pour chaque émission on prend les 22 chansons qui sont passées et 978 chansons qui ne sont pas passées.

Loss : multiclass pour essayer mais plutôt recall pour pénaliser les FN. On préfère apprendre des chansons FP que de ne pas apprendre des FN qui passent.

Idées :

* Ajouter une variable de clustering d’artiste
* Loss
* Retravailler le clustering des chansons
* Metrique : grid search pour les poids
* Etudier le rééquilibrage
* Ajouter des données