VINCENT GAGNON – GAGV25059800

Baccalauréat en Informatique

Rapport Projet Final

Travail d’apprentissage machine

présenté à

Julien Maitre

dans le cadre du cours

8INF436 Forage de données

Module d’informatique et de mathématique

Université du Québec à Chicoutimi

Le 4 Mai 2021

## Description de l’ensemble de données

L’ensemble de données choisi contient 42 305 chansons provenant de Spotify. Chaque chanson possède 22 attributs dont le titre et le genre. Nous chercherons à classer les chansons selon leur genre et donc entraînerons notre modèle pour qu’il puisse prédire le genre d’une chanson selon les caractéristiques de celle-ci. Voici les attributs de chaque instance :

### Danceability

Plus la valeur est élevée, plus il est facile de danser sur la chanson.

### Energy

Plus la valeur est élevée, plus la chanson est énergétique.

### Key

La clé de la chanson (Do, Ré, Mi, …). Prend les valeurs 0 à 11 qui correspondent aux douze notes que peut prendre la clé.

### Loudness

L’intensité sonore relative de la chanson en décibels.

### Mode

Valeur binaire représentant s’il s’agis d’une chanson commençant par une progression d’accords majeurs (1) ou non (0).

### Speechiness

Plus la valeur est élevée, plus la chanson contient des paroles.

### Acousticness

Plus la valeur est élevée, plus la chanson est acoustique

### Instrumentalness

Plus la valeur est proche de 1, plus la chanson est instrumentale.

### Liveness

Plus la valeur est élevée, plus il y a de chances que la chanson ait été enregistrée avec un public ou dans un concert.

### Valence

Plus la valeur est élevée, plus la chanson sonne positive.

### Tempo

Le tempo de la chanson, en battements par minute.

### Type

Type de contenu sur Spotify, ne prend qu’une seule valeur et n’est donc pas un attribut utile.

### Id

Identifiant unique de la chanson sur Spotify.

### Uri

Lien Spotify vers la chanson.

### Track\_href

Lien vers la chanson sur l’api de Spotify.

### Analysis\_url

Lien de l’analyse de la chanson sur l’api de Spotify.

### Duration\_ms

Durée de la chanson en millisecondes.

### Time\_signature

Mesure de la chanson.

### Genre

Genre de la chanson, prend 15 valeurs distinctes : Underground Rap (5875), Dark Trap (4578), Hiphop (3028), Trance (2999), Trap (2987), Techhouse (2975), DnB (2966), Psytrance (2961), Techno (2956), Hardstyle (2936), RnB (2099), Trap Metal (1956), Rap (1848), Emo (1680) et Pop (461).

### Song\_name

Nom de la chanson.

### Unnamed : 0

Numéro de la chanson dont le champ song\_name est null.

### Title

Titre des chansons dont le champ song\_name est null.

En voyant que l’ensemble de données ne contient que 42 305 chansons et 15 genres, il est facile de constater que celui-ci est loin d’être l’ensemble complet de toutes les chansons sur Spotify.

## Vérification et pré-traitement des données

L’ensemble de données contient des valeurs manquantes pour l’attribut song\_name, Unnamed : 0 et Title. Cependant, les instances donc la valeur est manquante pour song\_name, l’attribut Title est présent. On peut donc dire que l’ensemble de données ne contient pas de valeurs manquantes.

En regardant les diagrammes à boite des valeurs des différentes caractéristiques on s’aperçoit que certaines d’entre elles contiennent des valeurs extrêmes. En revanche, aucune d’entre elles ne sont des valeurs aberrantes et donc nous n’avons pas d’instances uniques à supprimer concernant les valeurs aberrantes.

Plusieurs attributs sont inutiles pour l’usage dont nous voulons en faire. Entre autres, tout ce qui a rapport aux titres des chansons et les identifiants et Uri Spotify, nous allons donc supprimer ces colonnes de l’ensemble de données. Nous supprimons aussi la colonne type car elle n’a qu’une seule valeur qui n’est pas significative.

Après avoir supprimer ces colonnes, on retrouve encore des valeurs catégoriques dans le dataset, entre autres les attributs key, mode et time\_signature. Je vais donc encoder ces attributs en utilisant le One Hot Encoding afin de les encodeurs. Cela consiste à créer une nouvelle colonne pour chaque catégorie d’un attribut et pour chaque instance on met un 1 dans la bonne colonne correspondant à la catégorie de l’attribut. En effectuant cela pour les colonnes key et time\_signature on obtient seulement des 0 et des 1 pour les attributs catégoriques du dataset.

On encode les caractéristiques catégoriques de cette façon car on ne veut pas débalancer les algorithmes que nous utiliserons plus tard dans ce travail, par exemple si on avait encoder les attributs en remplaçant les catégories par des chiffres, ceux-ci auraient pu débalancer l’algorithme car les catégories avec des nombres plus élevé auraient eu un plus grand poids que ceux ayant un nombre moins élevé. En mettant des 1 et des 0 on évite ce problème.

Après avoir encoder les attributs catégoriques, on standardise les valeurs des attributs numériques afin qu’une colonne n’ait pas des valeurs qui dominent les autres colonnes.