# Songs Genre

Yoann Ayoub et Louise Lizé

5 mars 2023

**Résumé**

Ce document présente une analyse du jeu de données "Songs Genre". Ce dernier provient de *Spotify* et son package *spotifyr*, créé par Charlie Thompson, Josiah Parry, Donal Phipps et Tom Wolff. Ce document présente les différentes analyses faites sur ce jeu de données et les résultats obtenus.

## 1 Présentation des données

Notre jeu de données est constitué de 23 variables et plus de 30 000 entrées. Il comporte environ 5000 pistes provenant de 6 genres de playlists : Edm, Latin, Pop, R&B, Rap et Rock.

Nous avons séparé les variables en quatres catégories :

1. Les variables qualitatives d'identification : pour les *pistes* avec trackID, trackName et trackArtist, pour les *albums* avec trackAlbumID, trackAlbumName et trackAlbumReleaseDate et pour les *playlists* avec playlistName, playlistID, playlistGenre et playlistSubgenre.
2. Les variables quantitatives de caractérisation :
   - *key* : représente les clés, 0 = C, 1 = C/D, 2 = D etc. -1 équivaut à une clé non détectée.
   - *mode* : 1 pour majeur et 0 pour mineur
   - *loudness* : niveau sonore global en décibels
   - *tempo* : le rythme en battements par minute
   - *durationMs* : la durée en millisecondes
3. Les variables quantitatives de descriptions des pistes (se basent sur des mesures et des algorithmes et prennent des valeurs entre 0 et 1) :
   - *danceability* : décrit si une piste est "dansable"
   - *energy* : donne la mesure d'intensité et d'activité de la piste. Par exemple, le métal se rapprochera de 1 et les préludes de Bach de 0.
   - *speechiness* : détecte la présence de mots "parlés" dans la piste. Par exemple, la poésie est proche de 1 et les genres très chantés comme le Rock de 0.
   - *acousticness* : comme son nom l'indique
   - *instrumentalness* : prédit si une piste est "non vocale". Les onomatopées "ooh" et "aah" sont traitées comme "instrumental" et le Rap ou les mots parlés comme "vocal".
   - *liveness* : détecte la présence d'un public dans l'enregistrement
   - *valence* : décrit la positivité transmise par une piste. 1 pour une piste joyeuse et 0 pour triste.
4. Enfin, la variable quantitative *popularity*, entre 0 et 100, représente la popularité d'une piste.

## 2 Analyse descriptive des données

### 2.1 Nettoyage des données

Les premières analyses ont permis de nettoyer et d'apporter des modifications à notre jeu de données. En étudiant la variable *duration*, on remarque que certaines pistes sont très courtes. Les pistes ayant une durée inférieure à 1 min ont donc été retirées. Pour faciliter nos analyses, nous avons également séparé la variable correspondant à la date en une variable année et une variable mois. Nous avons décidé d'utiliser l'année comme une variable quantitative par la suite car elle nous permet d'obtenir des résultats beaucoup plus intéressants. Cependant, il faudra penser à faire attention à l'usage de celle-ci dans certains cas, par exemple si nous voulons prédire des étiquettes sur de nouvelles musiques (après 2020). Nous avons aussi normé nos variables quantitatives à l'aide de la fonction *MinMaxScaler* de sklearn.

### 2.2 Analyse des données quantitatives

Dans un premier temps, nous avons effectué des boîtes à moustaches sur les différentes variables quantitatives. Nous avons pu constater des premières différences entre les genres. Nous pouvons prendre l'exemple de la variable *speechiness* (figure 1), où les valeurs prises par le genre Rap semblent plus dispersées et plus grandes que les autres genres. Cela semble cohérent

puisque cela signifie que c'est le genre qui a le plus de mots parlés.

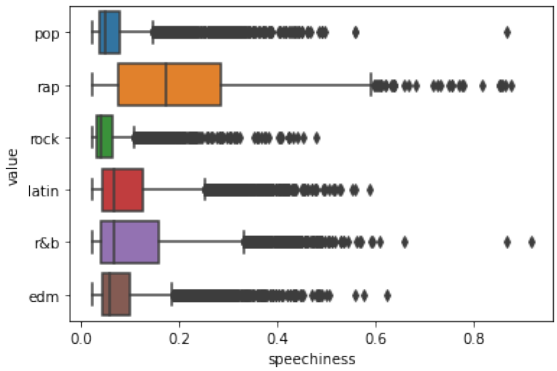


FIGURE 1 – Boxplot - speechiness en fonction du genre

Avec ce genre d'analyse, l'Edm nous semble être le genre le plus facile à distinguer des autres. Les boîtes à moustache le concernant sortent souvent du lot. On peut dire par exemple que la musique Edm semble avoir un *acousticness* plus faible que les autres genres. Elle a un tempo qui est très peu dispersé comparé aux autres et tourne autour de 125 battements par minute. Elle semble être la moins populaire et avoir une valence généralement plus faible que les autres. Les boîtes à moustache de *trackAlbumReleaseYear* sont également très différentes selon le genre. Les musiques Rock ont une très grande dispersion (1980-2010) contrairement aux autres qui commencent à partir des années 2000 (R&B, Rap) voire 2010 (Pop, Latin, Edm).

Nous avons essayé d'effectuer ces mêmes analyses selon la popularité des chansons. A part les chansons très populaires (>80/100) qui sont quasiment toutes situées autour de l'année 2019, nous n'avons pas remarqué de grandes différences des variables selon la popularité.

On peut également mettre en avant les premières corrélations entre nos variables grâce à la génération d'une heatmap. On observe un lien entre *loudness*, *energy* et *accousticness*. Ces corrélations pourraient se traduire ainsi : une musique dite énergétique a un fort niveau sonore et une musique acoustique semble avoir très peu d'énergie et peu d'intensité sonore, ce qui semble tout à fait cohérent. Pour poursuivre l'étude, nous avons analysé les corrélations au sein des différents genres. Cela nous a permis de distinguer les genres Rap et Edm, pour lesquels le nombre de variables corrélées entre elles est plus important et leur corrélation plus forte. Pour Edm, on retrouve notamment une corrélation entre popularité et instrumentalness-durée qui peut être intéressante.

## 2.3 Analyse des données qualitatives

On rappelle que le jeu de données est construit à partir de la récupération des musiques appartenant à des playlists. Ainsi, près de 25% des musiques de notre jeu de données se trouvent dans plusieurs playlist et apparaissent alors plus d'une fois.

Une variable qualitative qui a retenu notre attention est la variable *trackArtist*. Dans un premier temps, nous avons étudié le nombre de musiques présentes dans notre jeu de données par artiste. 50% des artistes ont plus de 6 musiques, ce sont les artistes "principaux". Les autres seront appelés les "petits" artistes. On obtient également une répartition de presque 50-50 du nombre de musiques produites par un gros ou un petit artiste. Le nombre de musiques par artiste présentes dans le jeu de données va jusqu'à 153 et environ 75% des artistes en ont moins de 17.

Nous nous sommes alors demandés si la popularité d'une musique était dépendante du fait qu'il s'agisse d'un artiste principal ou non. Nous avons alors utilisé le test *du CHI2 d'indépendance* vu en SY02. La p-value étant très proche de 0, nous pouvons rejeter l'hypothèse nulle : il semblerait alors que ces deux variables ne soient pas indépendantes. Cela se ressent surtout au niveau des musiques très populaires (>80/100), comme nous le montre également le tableau des fréquences attendues issu de ce test. Les petits artistes n'ont en réalité que 107 musiques très populaires, ce qui aurait dû se rapprocher de 707 et les artistes principaux en ont 1233 alors que le tableau des fréquences donne un résultat de 633. Nous avons recommencé en remplaçant la variable de popularité par les genres mais nous ne percevons pas de résultats aussi intéressants.

Nous nous sommes également demandés si nos artistes principaux avaient un genre de musique de prédilection. Les 864 artistes principaux ont une moyenne de 2.25 genres. Un tiers d'entre eux en ont même qu'un seul. Ainsi, nous nous demandons si prendre en compte l'artiste et le style de son répertoire pourrait nous aider dans la prédiction du genre d'une musique.

Nous pensions également trouver des différences de la variable *trackAlbumReleaseMonth* selon les genres, comme par exemple plus de musiques latines l'été mais il ne semble pas vraiment y avoir d'impact.

Ensuite, nous avons voulu étudier la variable *playlistName* dans le but de pouvoir identifier le genre d'une musique selon le titre de sa playlist. C'est cela qui nous a permis de comprendre que la variable *playlistGenre* était, comme son nom l'indique, le genre de la playlist dans laquelle appartient la musique et non pas directement le genre de la musique. Nous avons alors remarqué

qu'une partie de nos musiques appartenait à plusieurs genres de playlist à la fois. Nous avons sûrement été influencés par le titre "Song genres" ou en lisant l'article mentionné sur le gitlab du tidyTuesday qui n'éclaircit pas cette petite ambiguïté [2].

### 2.4 Conclusion de l'analyse exploratoire

Cette partie nous a ainsi aidé à définir nos axes d'analyse. Nous étudierons tout d'abord le genre d'une musique à partir de la variable *playlist_genre* donnant le genre de la playlist dans laquelle elle se trouve. Puis dans un deuxième temps, nous chercherons à prédire le genre d'une playlist selon les musiques qui la composent.

## 3 Genre (de la playlist) d'une musique

Il est tout d'abord important de noter que dans cette partie, quand nous parlons du genre d'une musique nous parlons en réalité du genre de la playlist dans lequel se trouve la musique.

Dans un premier temps, nous allons présenter l'application des méthodes non supervisées sur notre jeu de données. La directive de notre recherche ne se prête pas vraiment à ces méthodes mais nous avons tout de même souhaité l'effectuer pour essayer d'en apprendre plus sur notre jeu de données. Dans un second temps, nous étudierons les méthodes supervisées.

### 3.1 Méthodes non supervisées

#### 3.1.1 Analyse Composantes Principales

L'ACP sur notre jeu de données nous permettrait de réduire le nombre de variables quantitatives en 6 axes principaux d'inertie, représentant un peu plus de 80% de l'information. La représentation de nos données selon ces axes ne nous a pas permis de différencier les musiques selon leur genre. Cependant, cela a pu aider à mieux comprendre la corrélation entre nos variables. Le cercle de corrélation selon l'axe 1 et 2, conservant presque de 40% d'informations se trouve en figure 2. Les principales observations que l'on peut faire dans ce premier plan sont que *trackPopularity* contribue beaucoup au premier axe, *acousticness* est plutôt anticorrélée à *valence* et à *energy*, et *instrumentalness* et *energy* sont également assez anticorrélées. Cela semble cohérent avec ce qui avait été dit en analyse exploratoire.

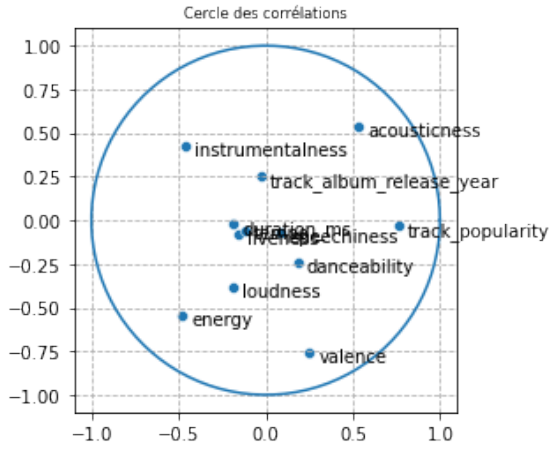


FIGURE 2 – Cercle de corrélation - 1er plan factoriel

#### 3.1.2 Algorithme de clustering

Ayant un grand volume de données, la CAH ne semble pas très adaptée à notre étude. L'algorithme des K-MEANS a donc plutôt été privilégié. Le nombre de clusters choisi est donc 6 pour essayer de mettre en avant les 6 genres des musiques. Nous n'avons malheureusement pas de nette séparation. Mais certaines données nous montrent tout de même des différences entre les genres. Par exemple, voici un des résultats retournés par le KMeans : 57% des musiques du cluster 1 sont Edm, 57% de celles du cluster 3 sont de type Rock et ce même groupe contient presque 0% de musiques Edm. Cela semble rejoindre nos observations de départ, disant que Edm et Rock semblaient être plus faciles à séparer des autres genres.

### 3.2 Méthodes supervisées

Dans cette partie, les étiquettes choisies pour les différentes méthodes seront donc nos 6 genres de musique. Nous rappelons que notre jeu de données possède une répartition plutôt bien équilibrée selon les genres. Notre but étant de prédire les genres, nous utiliserons des méthodes de classifications.

#### 3.2.1 Méthode des K plus proches voisins

Cette méthode nous permet d'obtenir un score de 0.51, qui pourrait sembler être un résultat plutôt moyen. Il faut prendre en compte l'avertissement de scikit qui nous rappelle que dans le cas multiclasse, le score peut s'avérer parfois "dur". En effet, cela signifie tout de même que l'on passe d'environ 1 chance sur 6 de trouver

le genre d'une musique à 1 chance sur 2. Le score ainsi obtenu nous servira de score référence que nous chercherons à améliorer par la suite. Pour essayer d'avoir un résultat plus concluant, nous avons réduit le nombre d'étiquettes à 2, en confrontant à chaque fois deux genres de musique. Une confrontation avec Rock donne en moyenne de très bons résultats (jusqu'à 0.91), mais aussi Edm et Rap comme nous l'avions prévu. Sur l'ensemble des confrontations à 2, on trouve une moyenne des scores supérieure à 0.80, ce qui est plutôt satisfaisant. Nous avons également essayé une confrontation à trois genres entre Edm, Rap et Rock, qui nous donne un score moyen encore satisfaisant de 0.81. Ainsi qu'une confrontation à 4 genres, en y ajoutant Latin, pour lequel nous obtenons un score de 0.7.

### 3.2.2 Autres classifieurs : LDA, QDA, bayésien naïf, 1-VS-1, 1-VS-REST

Nous avons ensuite essayé d'autres classifieurs en espérant améliorer le score obtenu avec les KNN. Dans un premier temps, nous avons voulu vérifier que nos classes suivaient une loi normale multidimensionnelle. En effectuant en test de Shapiro-Wilk sur chaque classe, l'hypothèse nulle était rejetée. Il est donc très peu probable d'avoir de telles données en supposant qu'elles soient normalement distribuées. Après quelques recherches, il semblerait qu'avoir un grand jeu de données permet tout de même d'obtenir des résultats concluants.

Nous avons répété une dizaine de fois les différents classifieurs. Ils nous donnent des résultats très similaires à ceux obtenus avec les KNN, avec un score proche de 0.5 sur les 6 genres et des scores en moyenne similaire en confrontant les genres à 2.

### 3.2.3 Arbre de décision et Random Forest

Nous avons ensuite utilisé la méthode des arbres de décisions binaires sur notre jeu de données. Le score moyen obtenu avec cette méthode pour la classification des 6 genres n'est pas très bon : 0.45 si on construit l'arbre en entier et 0.4 si on se limite aux 6 noeuds de la figure 3. De plus, le critère d'impureté de chaque classe de cette figure est élevé, avec un gini compris entre 0.5 et 0.8. Ainsi, bien que le résultat soit loin d'être satisfaisant, l'intérêt principal de cette méthode est de montrer les variables qui sont les plus importantes dans la classification. La figure 3 présente l'arbre le plus simple possible avec chacun des genres majoritaires dans un groupe. A noter qu'il a fallu forcer une telle construction avec les arguments *max_ depth* et *min_ samples_ leaf* du DecisionTreeClassifier de sklearn et que cette figure s'appuie sur le package dtreeviz [1].

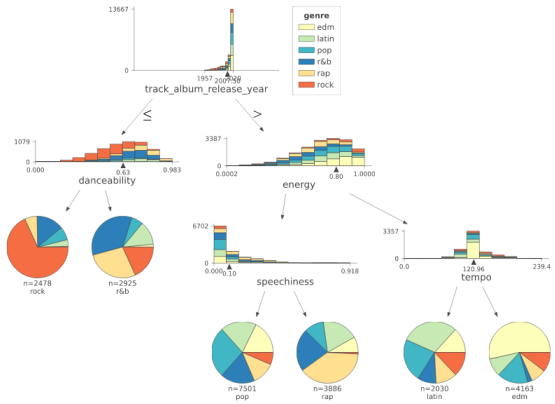


FIGURE 3 – Arbre de décision

On observe que la donnée la plus importante pour déterminer le genre d'une musique est son année de sortie avec une *feature_ importance_* égale à 0.36. On remarque que certains genres sont plus faciles à caractériser que d'autres : Rock par une ancienne année de sortie et une faible danceability, Rap par un speechness élevé et Edm par un tempo élevé. Les trois autres genres musicaux semblent plus difficiles à classifier puisqu'ils sont plus répartis entre les différentes classes.

Afin d'utiliser au mieux les arbres binaires et dans l'optique d'obtenir le meilleur score possible, nous avons appliqué la méthode Random Forest sur notre jeu de données. On obtient un score de 0.57 sur la classification des 6 genres avec cette méthode. Ceci représente le meilleur score obtenu sur le jeu de données contenant l'ensemble des 32 000 musiques.

### 3.2.4 Étude sans les morceaux ambigus

Il existe des musiques qui apparaissent plusieurs fois dans différentes playlists. Pour certaines, l'ensemble de ces playlists sont du même genre. Pour d'autres, les playlists sont de genres différents (pour la même musique). Nous considérerons donc que si une même musique apparaît dans des playlists de genres différents alors le genre de cette musique est ambigu. Nous avons donc construit un nouveau jeu de données en retirant l'ensemble des musiques avec un genre ambigüe. En appliquant les différentes méthodes supervisés sur les 23 000 données ainsi obtenues, on observe une amélioration nette du score pour chacune d'elle. Pour la classification des 6 genres, cette amélioration est comprise entre 5% pour les KNN avec un score de 0.56, et 7% pour le random forest avec un score de 0.64.

#### 3.2.5 Étude sur les artistes principaux

Comme évoqué en analyse exploratoire, les musiques d'un même artiste semblent appartenir à 2 ou 3 genres de playlists en moyenne. Nous avons essayé d'utiliser cette information dans nos méthodes supervisées. Nous avons donc créé un nouveau jeu de données en ne gardant que les 13 000 morceaux appartenant aux artistes principaux ($>$ 6 morceaux dans le jeu de données). En appliquant les différents classifieurs sans l'information des genres dans le répertoire de l'artiste, on trouve un score moyen de 0.55. En prenant en compte cette information, on augmente le score à 0.77. Pour ce faire, nous avons premièrement diviser le jeu de données avec 2/3 de données d'entraînement et 1/3 de données de test. Puis nous avons construit un tableau des fréquences de genres selon l'artiste à partir des données d'entraînement. Et pour finir, nous avons ajouté les 6 colonnes de fréquence aux données d'entraînement et de test.

# 4 Analyse du genre des playlists

Nous avons appliqué le même cheminement de réflexions que la précédente partie. Pour pouvoir analyser le genre de la playlist à partir des musiques qui la constitue, nous avons essayé pour chaque variable de faire une moyenne pour toutes ses musiques.Nous avons également essayé de prendre en compte la variance mais cela ne donnait pas de résultats concluants. Nous nous concentrons par la suite uniquement sur la moyenne par variable. Ainsi, nous avons réduit nos données à 449 entrées correspondant à chaque playlist.

## 4.1 Méthodes non supervisées

### 4.1.1 Analyse Composantes Principales

L'ACP nous permet de résumer 80% de l'information avec 5 axes principaux d'inertie. Le premier plan ne représente que 50% de l'information. On peut tout de même remarquer une petite distinction selon les genres dans ce premier plan (cf. figure 4).

### 4.1.2 Algorithme de clustering

En moyenne, l'application des K-means nous permet d'obtenir un score de rand de 0.26, ce qui n'est pas vraiment bon. En s'y intéressant de plus près, certains résultats semblent tout de même plutôt satisfaisants. Prenons un des résultats que KMeans nous a retourné : 70% des playlists *latin* ont été classées dans le cluster 0, 70% des playlists *rock* dans le cluster 3 et on retrouve

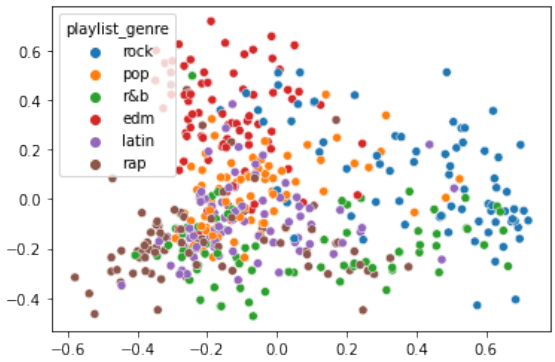


FIGURE 4 – Représentation ACP dans le premier plan

même 80% des playlists *edm* dans le cluster 5. Nous nous serions attendus à un résultat similaire pour le *rap*, mais l'algorithme a principalement classé les playlists de ce genre dans deux clusters distincts (50% dans le cluster 4 et 35% dans le 0). Nous avons même plusieurs genres qui n'apparaissent jamais dans certains clusters.

## 4.2 Méthodes supervisées

### 4.2.1 Classification

Nous avons récupéré 50 fois les scores des méthodes selon les différents classifieurs en figure 5. On trouve un score qui tend vers 0.70.

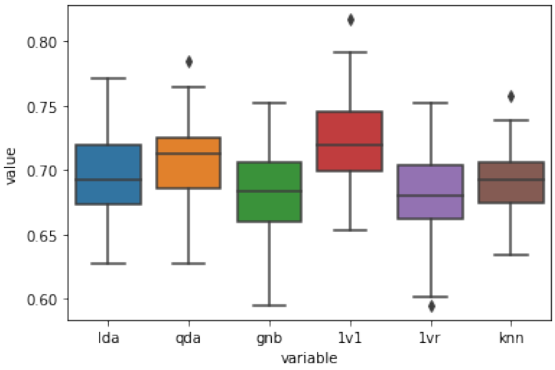


FIGURE 5 – Comparaison des différents classifieurs

Nous avons ainsi généré les matrices de confusion correspondantes. Nous pouvons voir un exemple d'une des meilleures matrices obtenue pour le classifieur One VS One figure 6. En voyant les scores aussi élevés sur la diagonale et aussi proches de zéro ailleurs, nous sommes très satisfaits des résultats.

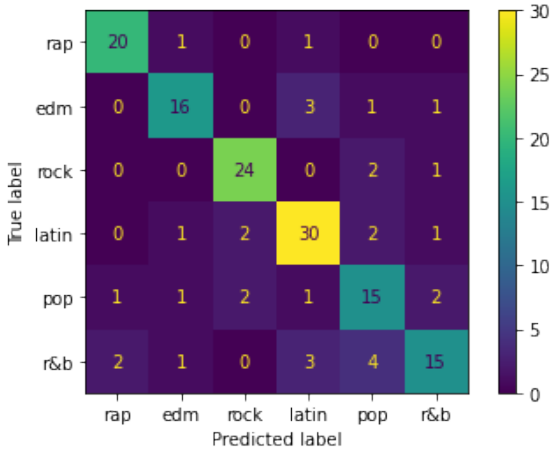


FIGURE 6 – Matrice de confusion 1 VS 1

### 4.2.2 Arbre de décision sur les playlists

Enfin, nous avons appliqué les méthodes se basant sur les arbres de décision en procédant de la même façon que pour la partie 3.2.3. Encore une fois, nous avons un score moyen inférieur à celui des classifieurs précédents : 0.65 pour les 6 genres avec l'arbre en entier et 0.6 en se limitant aux 6 premiers noeuds de la figure 7. Ces scores sont cependant bien meilleurs qu'avec le jeu de données des morceaux et le critère d'impureté le montre, avec un gini compris entre 0.3 et 0.6 pour les différentes classes. On remarque également que l'on a des classes contenant particulièrement un même genre pour le Rock (80%), Latin (78%) et Rap (76%). A noter que cet arbre a été plus facile à élaborer que le précédent puisqu'il est construit naturellement par l'algorithme quand on ne garde que les 6 premiers noeuds.

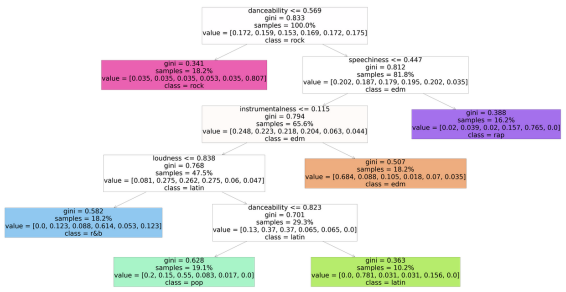


FIGURE 7 – Arbre de décision sur les playlists

Il est également intéressant de noter que ce ne sont pas les mêmes variables qui sont utilisées quand on applique cette méthode aux playlists ou aux morceaux. Ici, l'année de sortie, le tempo et l'énergie sont remplacés par les variables *loudness* et *instrumentalness*. On peut alors essayer de résumer les caractéristiques principales des playlists de chaque genre dans la table 1.

TABLE 1 – Caractéristique principale de chaque genre

| Genre | Caractéristique |
|---|---|
| Rock | Difficile de danser dessus |
| Rap | Beaucoup de paroles |
| Edm | Très instrumental |
| R&B | Un son faible |
| Latin | Très facile de danser dessus |
| Pop | Tout le reste |

Enfin nous avons appliqué la méthode Random Forest avec laquelle nous avons obtenu un score de 0.74. Cette méthode est également la plus performante pour ce jeu de données puisque le score obtenu est légèrement supérieur à celui obtenu avec les autres classifieurs.

## 5 Conclusion

Il n'y a pas vraiment de réponse binaire aux questions de prédictions que nous nous sommes posées au début, mais on peut dire que les différentes méthodes utilisées donnent toujours de meilleurs résultats que si l'on faisait des prédictions au hasard. Certaines caractéristiques que l'on a découvertes en analyse exploratoire nous ont permis d'effectuer de la prédiction sur des jeux de données plus petits selon certains critères. Avec ces modifications, on obtient de très bon scores de prédiction, surtout si l'on prend en compte le fait que l'on cherche à classer les données dans 6 étiquettes.

Ce jeu de données était très intéressant à analyser et nous pensons avoir beaucoup appris de lui. Il reste certainement plein d'autres connaissances à en tirer. Une autre directive aurait été de se demander si on pouvait prédire la popularité d'une piste.

## Références

[1] parrt. dtreeviz : Decision tree visualization. Technical report.

[2] K. Pavlik. Understanding + classifying genres using spotify audio features. Technical report.