IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Analyse des données de santé : épidémiologie et aide à la décision

PROJET : Effet de la musique sur la santé mentale

BEN JEMAA Yosr, DRIRA Yosr

SOMMAIRE

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

I - EXPLORATION DES DONNÉES

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

EXPLORATION DES DONNÉES

1.1 Contexte

- La thérapie musicale utilise la musique pour améliorer le stress, l'humeur et la santé mentale.

- Le jeu de données cherche à trouver des liens entre les préférences musicales et la santé mentale.

- Les résultats pourraient guider l'application de la thérapie musicale ou offrir des perspectives sur le fonctionnement de l'esprit.

Jeu de données: 700 responses (Music x Mental Health Traits) - 33 columns

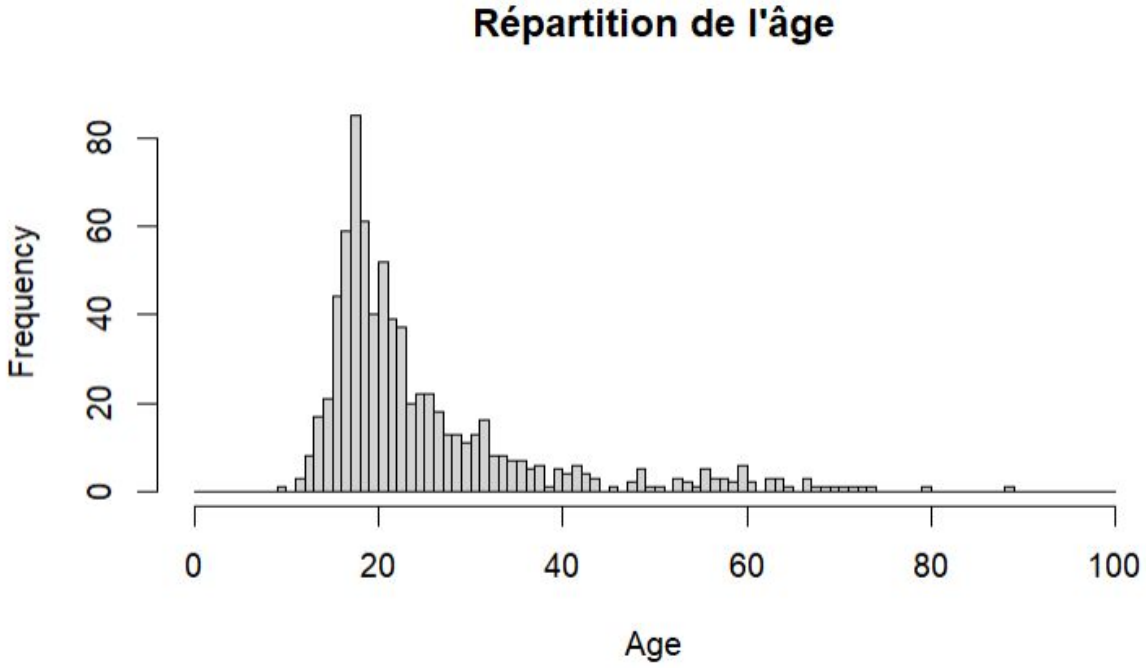
Variables Quantitatives	Variables Qualitatives
Timestamp	Primary streaming service
Age	While working
Hours per day	Instrumentalist
BPM	Composer
Frequency [genre]	Fav genre
Anxiety/ Depression/ Insomnia/ OCD	Exploratory
	Foreign Language
	Music effects

Source: https://www.kaggle.com/datasets/catherinerasgaitis/mxmh-survey-results

Comment la préférence musicale et les habitudes d'écoute peuvent-elles influencer les niveaux d'anxiété, de dépression, d'insomnie et les troubles obsessionnels-compulsifs (TOC) chez les individus ?

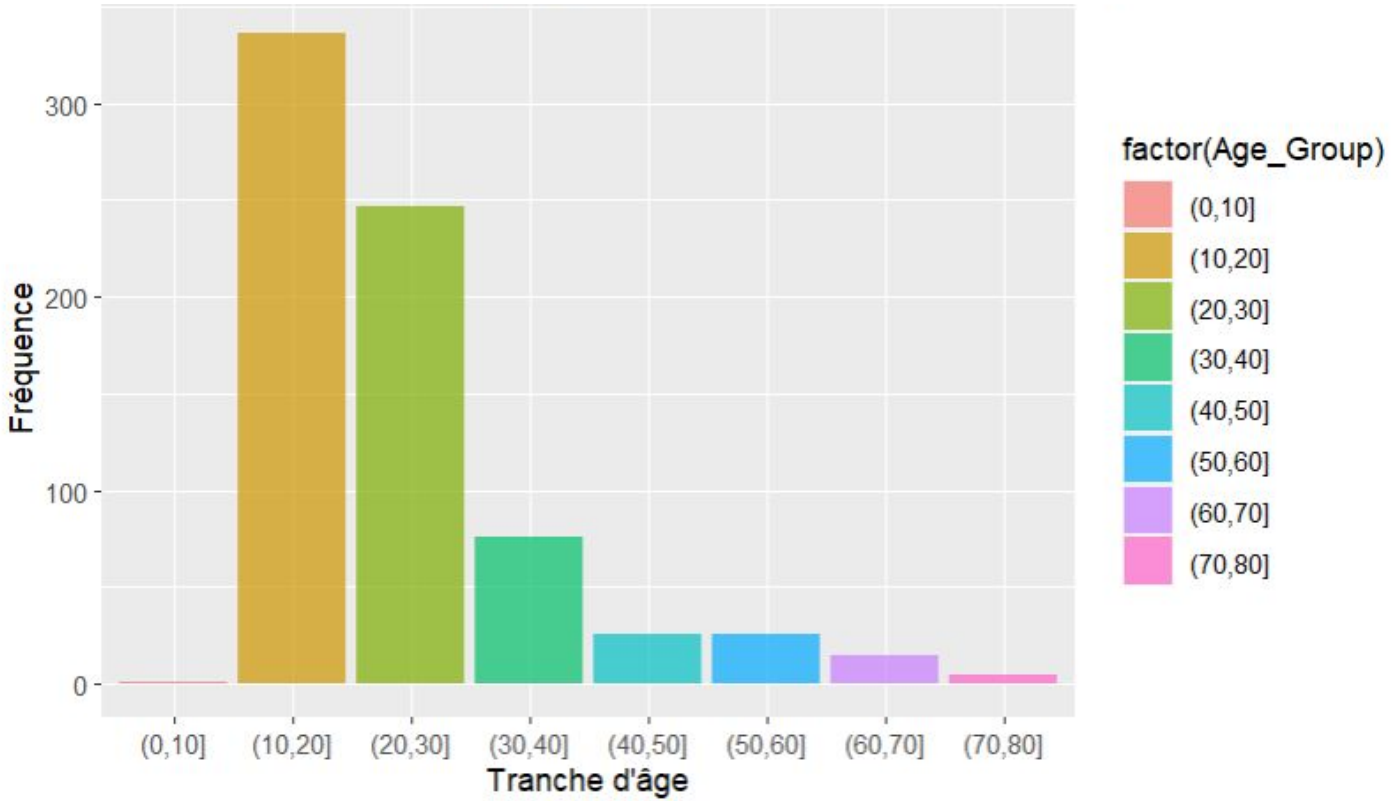
EXPLORATION DES DONNÉES

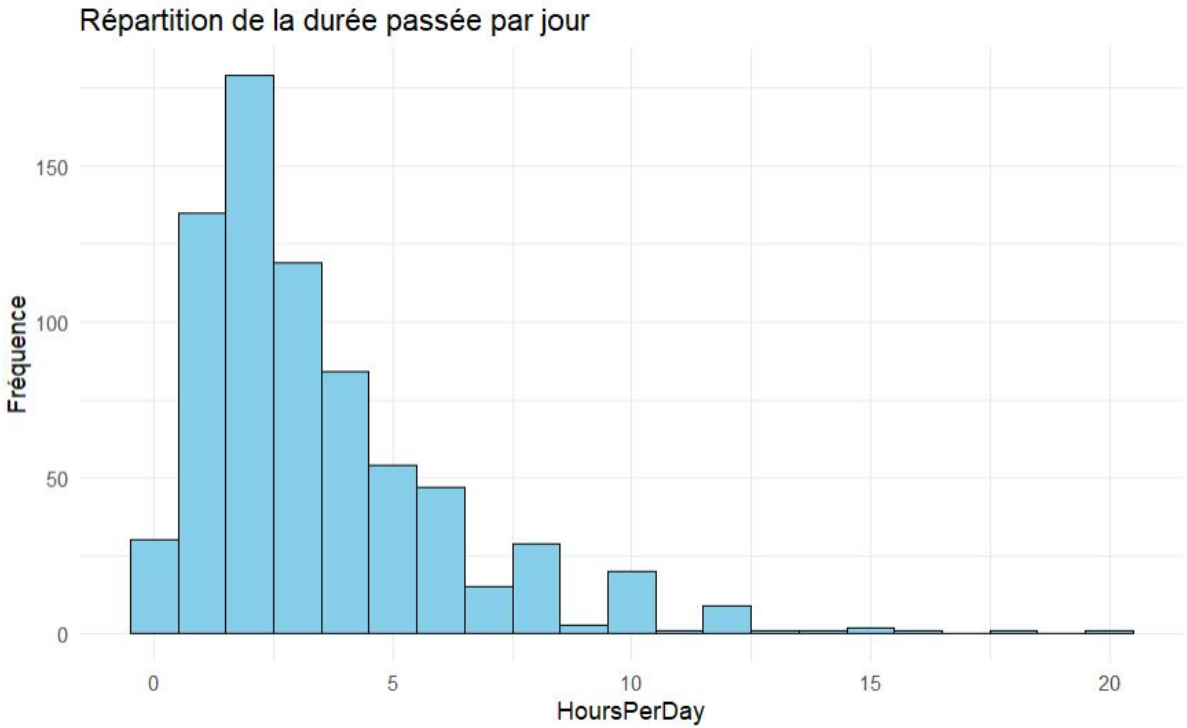
1.3 Démographie & comportements



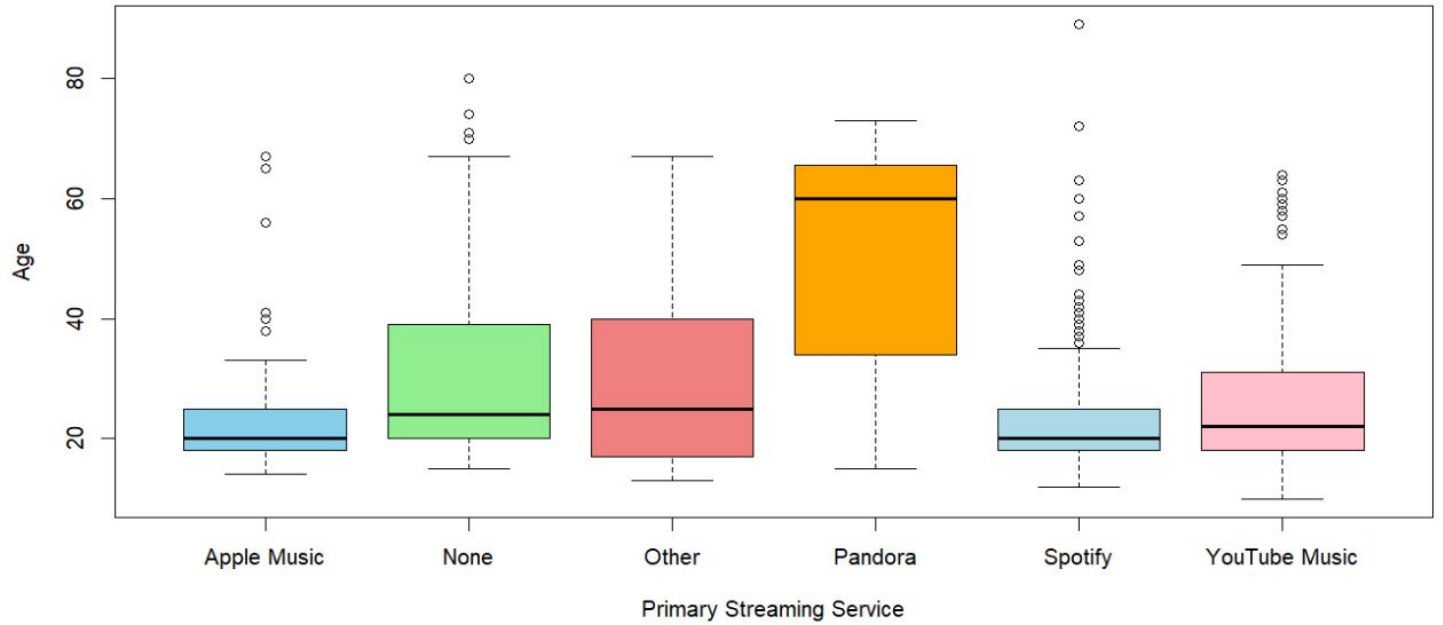
EXPLORATION DES DONNÉES

1.3 Démographie & comportements

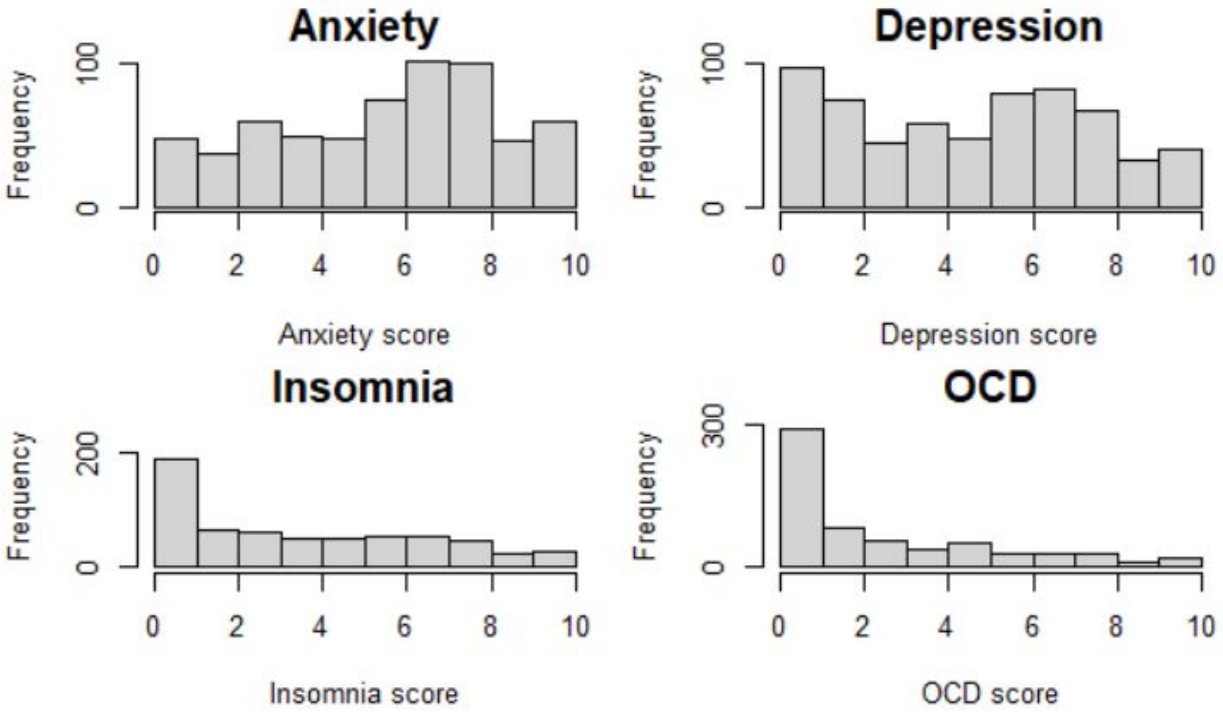



Répartition de la durée passée par jour
Fréquence
150
100
50
0
0
5
10
15
20
HoursPerDay

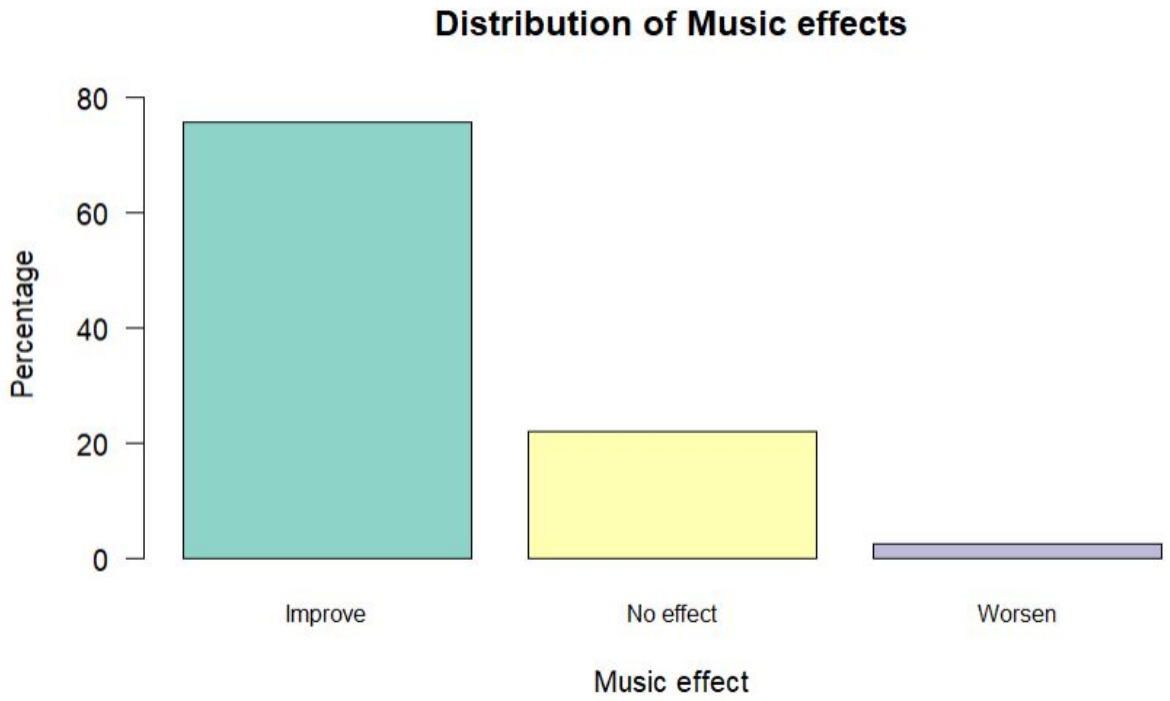
Diagrammes en BoxPlot de l'âge en fonction du service de streaming



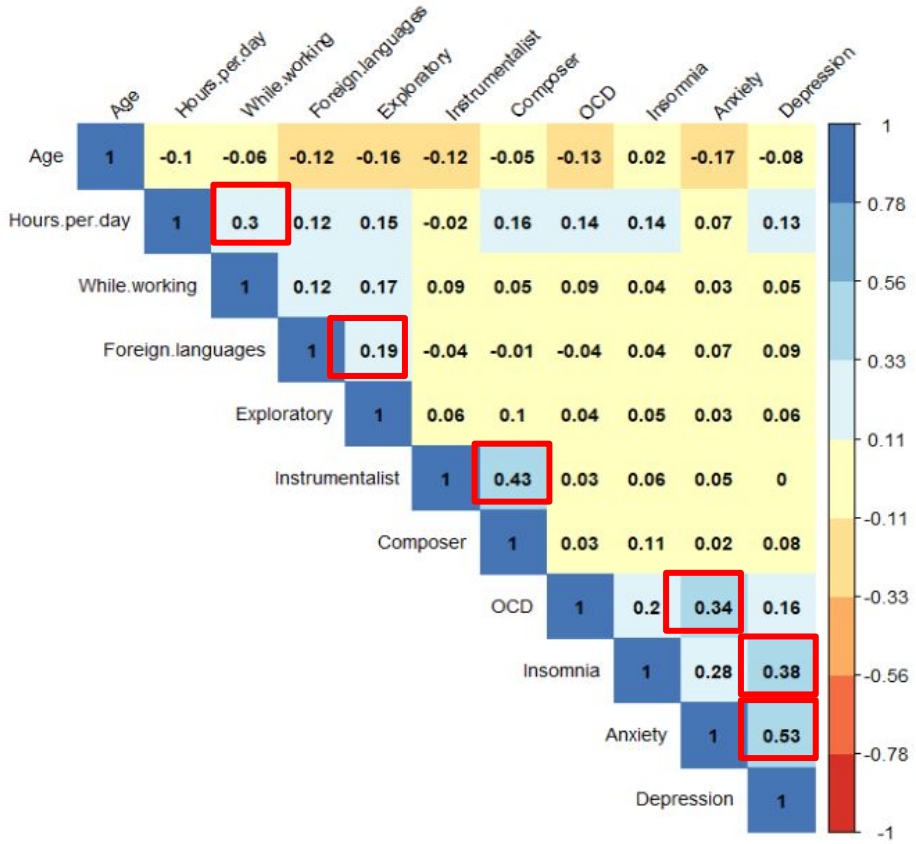
Histogrammes de répartition de fréquence des degré des indicateurs



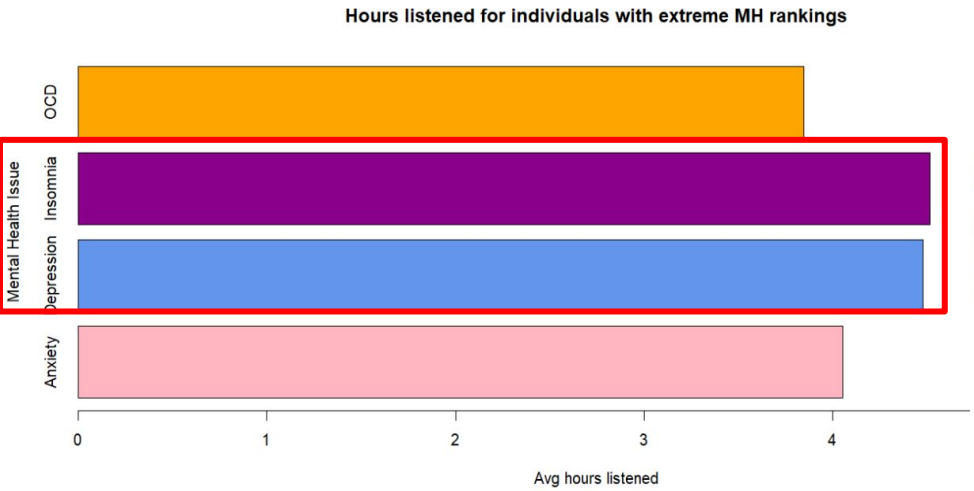
Distribution de la répartition de l'effet de la musique

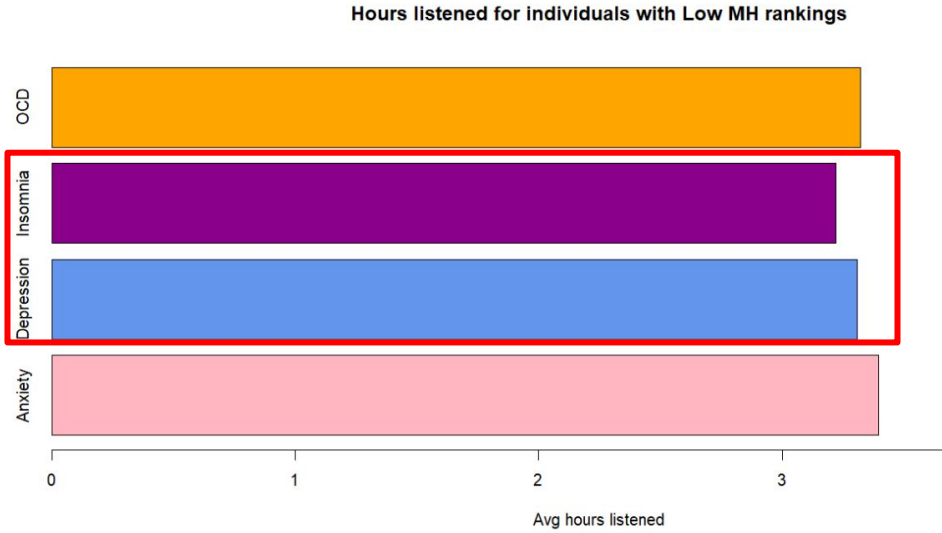


Heatmap


Age
Hours.per.day
While.working
Foreign.languages
Exploratory
Instrumentalist
Composer
OCD
Insomnia
Anxiety
Depression
1 -0.1 -0.06 -0.12 -0.16 -0.12 -0.05 -0.13 0.02 -0.17 -0.08
1 0.3 0.12 0.15 -0.02 0.16 0.14 0.14 0.07 0.13
1 0.12 0.17 0.09 0.05 0.09 0.04 0.03 0.05
1 0.19 -0.04 -0.01 -0.04 0.04 0.07 0.09
1 0.06 0.1 0.04 0.05 0.03 0.06
1 0.43 0.03 0.06 0.05 0
1 0.03 0.11 0.02 0.08
1 0.2 0.34 0.16
1 0.28 0.38
1 0.53
1
1
0.78
0.56
0.33
0.11
-0.11
-0.33
-0.56
-0.78
-1

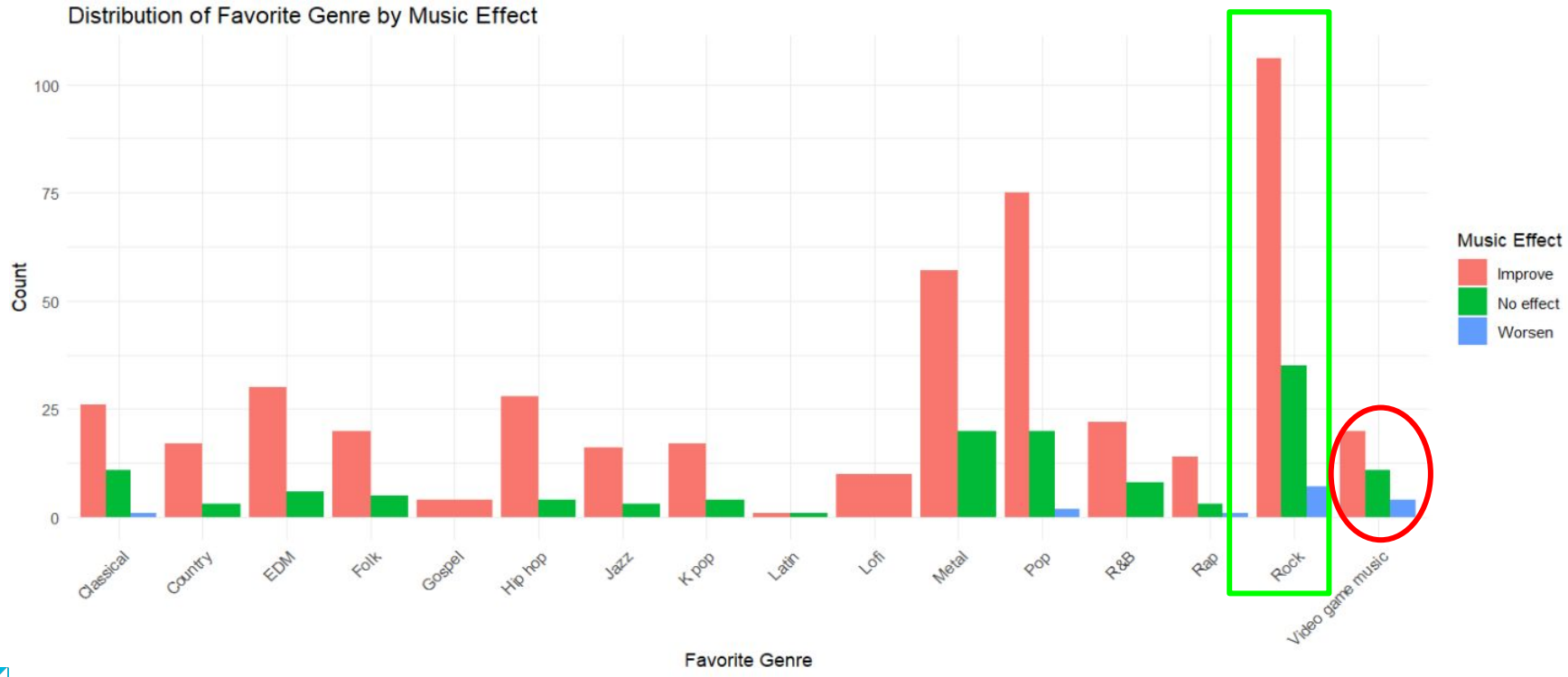
Impact du nombre d'heures d'écoute de musique sur les indicateurs

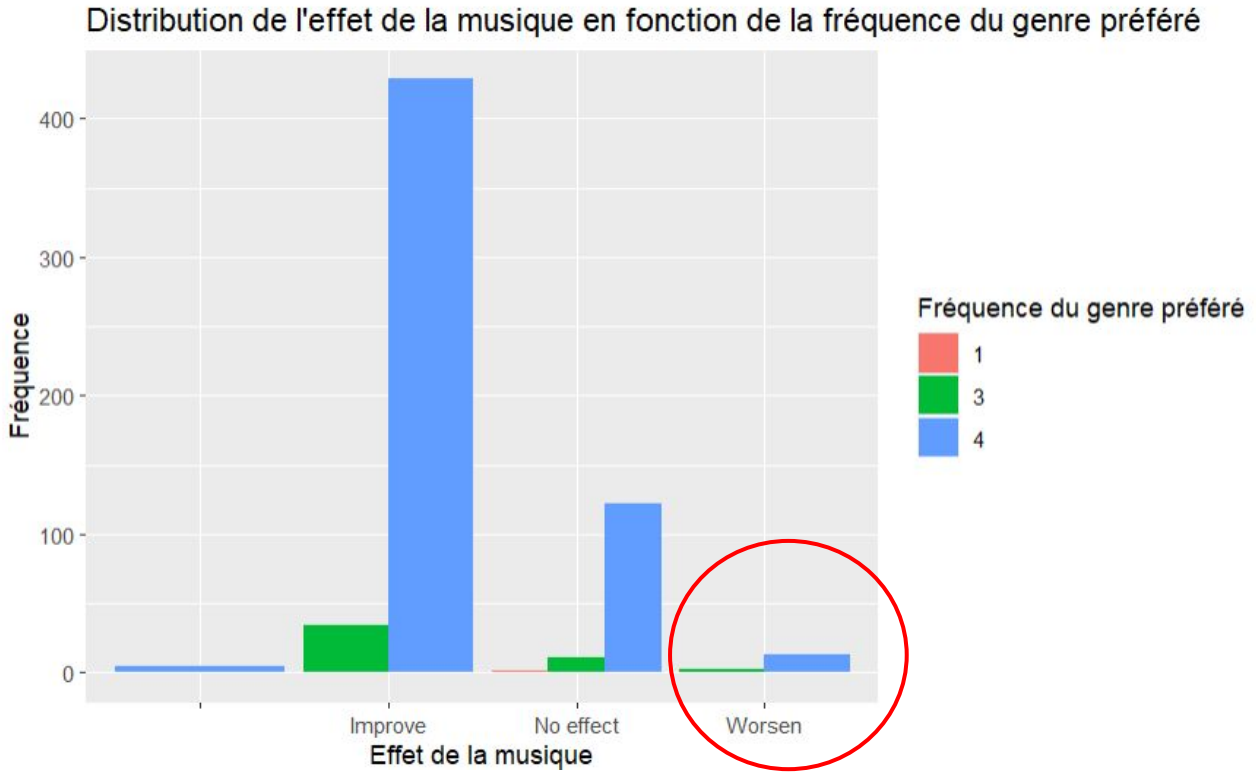


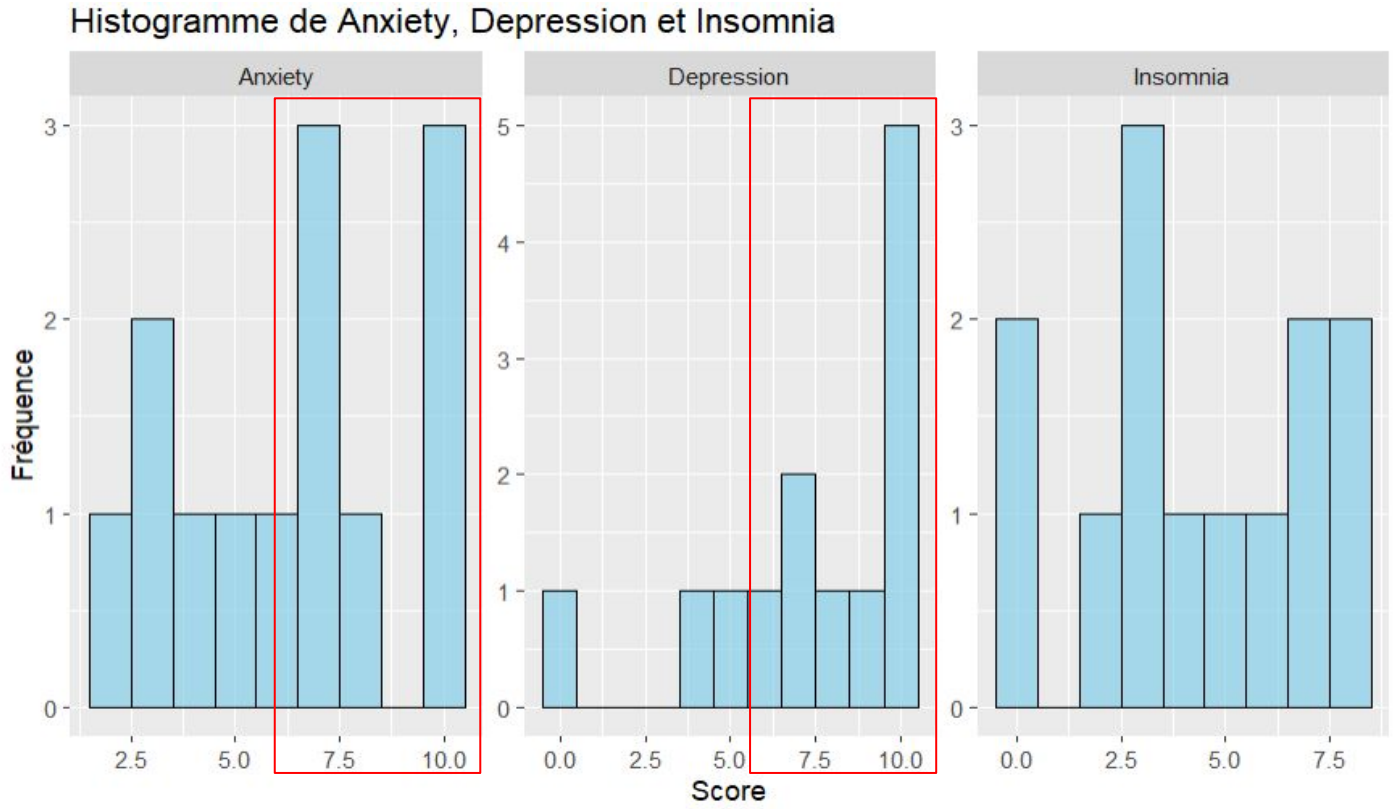


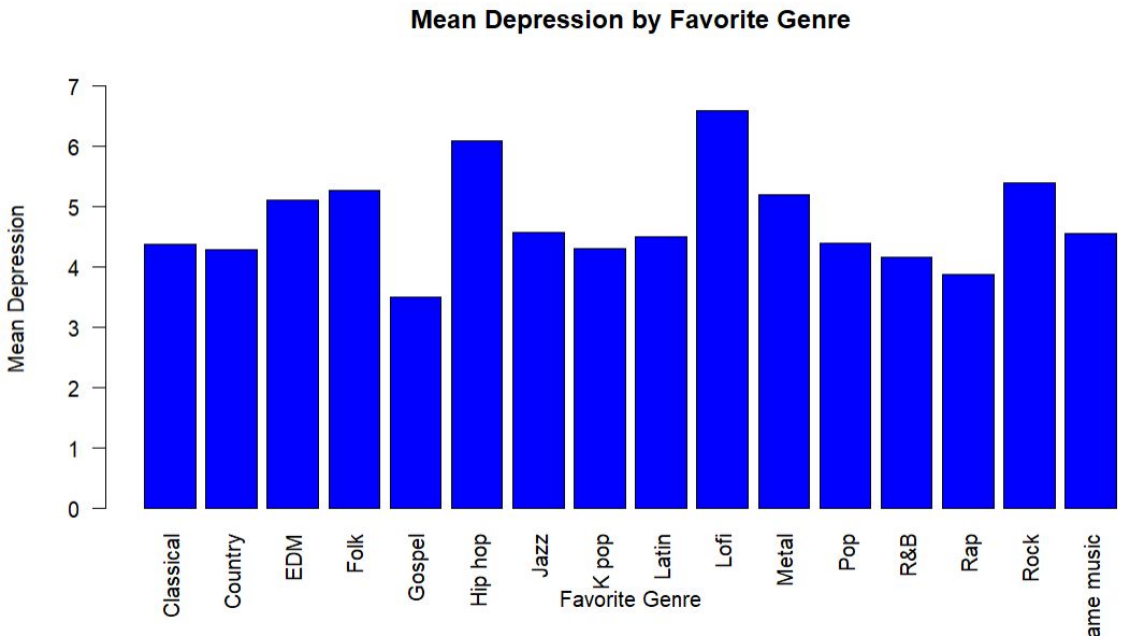
EXPLORATION DES DONNÉES

1.5 Exploration des relations

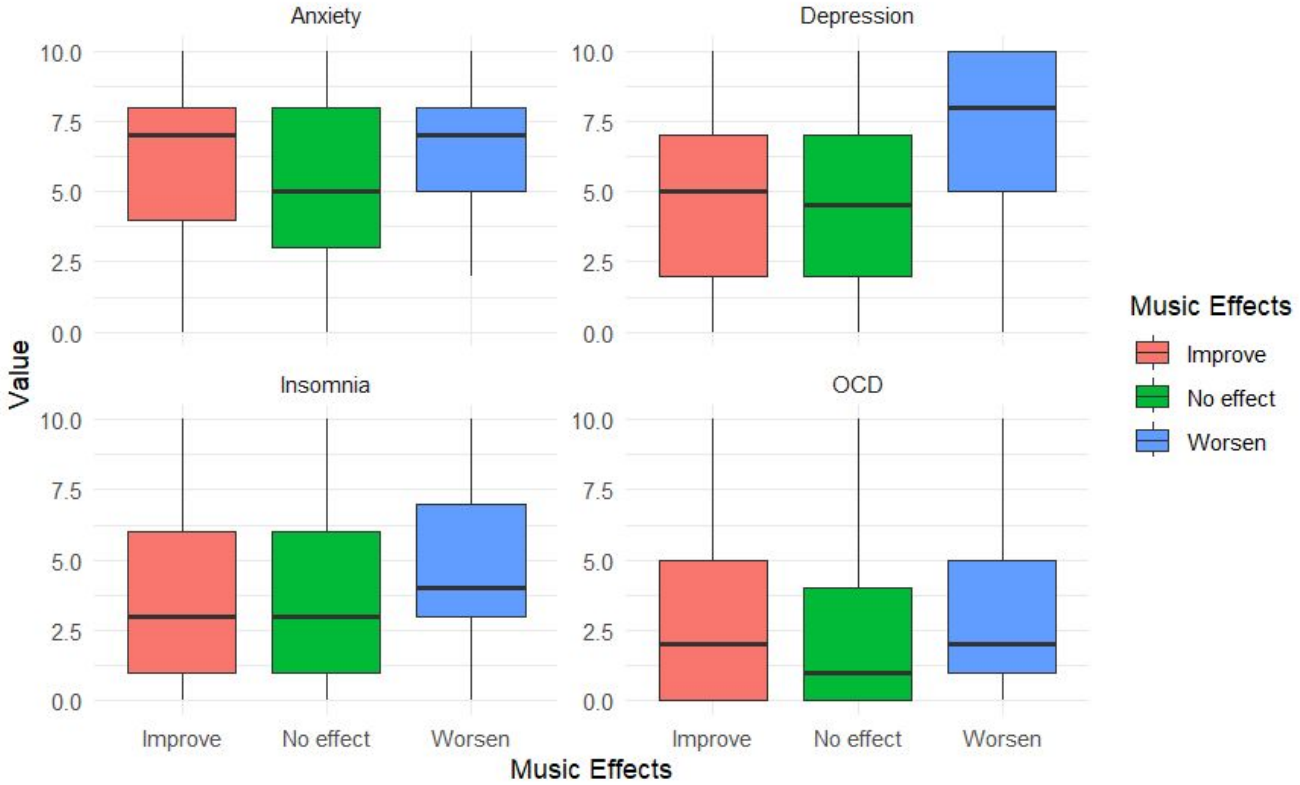



Distribution de l'effet de la musique en fonction de la fréquence du genre préféré
Fréquence
400
300
200
100
0
Improve
No effect
Worsen
Effet de la musique
Fréquence du genre préféré
1
3
4


Histogramme de Anxiety, Depression et Insomnia
Anxiety
Depression
Insomnia
Fréquence
Score
0
1
2
3
4
5
2.5
5.0
7.5
10.0
0.0


Mean Depression by Favorite Genre
Mean Depression
0
1
2
3
4
5
6
7
Classical
Country
EDM
Folk
Gospel
Hip hop
Jazz
K pop
Latin
Lofi
Metal
Pop
R&B
Rap
Rock
ame music
Favorite Genre

Boxplots des mesures de santé mentale selon les effets de la musique

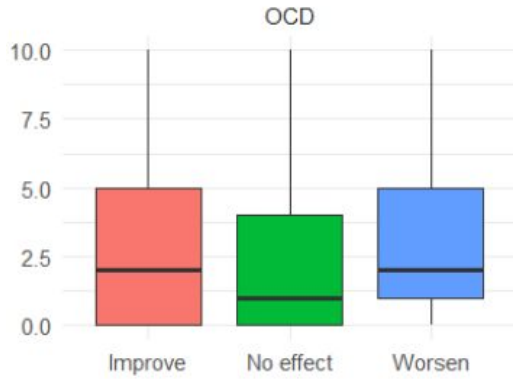


Relation entre les Indicateurs de santé mentale et effet de la musique

Le test de **Kruskal-Wallis** est une méthode non paramétrique utilisée pour comparer les valeurs médianes de trois groupes indépendants ou plus.

Hypothèse nulle (H0) :

L'OCD n'a pas d'impact significatif sur la perception de l'effet de la musique



Résultats du test Kruskal-Wallis Test

```
data:  Music.effects by OCD
Kruskal-Wallis chi-squared = 15.329, df = 12, p-value = 0.2239

        Kruskal-Wallis rank sum test

data:  Music.effects by Anxiety
Kruskal-Wallis chi-squared = 16.879, df = 11, p-value = 0.1115

        Kruskal-Wallis rank sum test

data:  Music.effects by Depression
Kruskal-Wallis chi-squared = 30.829, df = 11, p-value = 0.001173
```

→ Le COD / l'anxiété n'ont pas un impact significatif sur la perception de l'effet de la musique.
→ La dépression et l'effet de la musique sont liés.

II. ANALYSES SUPERVISÉES

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Régression

Objectif : Prédire les scores de santé mentale (MH Rankings) en fonction des caractéristiques disponibles.

Approches de modélisation :

Régression linéaire

Random Forest

XGBoost.

ANALYSES SUPERVISÉES

- **Entraînement :** Nous avons entraîné le modèle en utilisant 80 % de l'ensemble de données, en réservant les 20 % restants pour les tests.
- **Évaluation :**

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are predicted values

$y_1, y_2, \ldots, y_n$ are observed values

n is the number of observations

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i|$$

Where:

$\hat{y}_i$ = Predicted value for the i^{th} data point

y_i = Actual value for the i^{th} data point

n = number of observations

Evaluation des modèles

Target = Dépression

Métrique	Linear Regression	Random Forest	XGBoost
RMSE	3.09	3.08	3.16
MAE	2.62	2.64	2.65

Méthodes d'Optimisation :

- **Recherche de Grille** : Exploration d'un ensemble prédéfini de hyperparamètres pour trouver la combinaison optimale.
- **Validation Croisée :** Évaluation des performances du modèle sur plusieurs sous-ensembles des données pour estimer sa capacité de généralisation

```
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were nrounds = 50, max_depth = 3, eta =
0.1, gamma = 0, colsample_bytree =
 1, min_child_weight = 1 and subsample = 1.
XGBoost Model Evaluation for Depression:
RMSE: 3.129314
MAE: 2.680435
```

⇒ Il n'y a pas une grande différence

Normalization de la target: Score de dépression

Métrique	Random Forest	XGBoost
RMSE	0.30	0.34
MAE	0.26	0.28

Classification

Objectif :

Prédire l'effet de la musique à partir des indicateurs (Dépression, Anxiété, Insomnie, TOC/OCD).

Approches de modélisation :

Arbre de décision

Random Forest

Multinomial Logistic Regression

Prédiction de l'effet de la musique sur la santé mentale

1. Arbre de décision

Modèle : Arbre de décision

Bibliothèque utilisée: caret

tree_model <- rpart(Music.effects ~ OCD + Anxiety + Depression + Insomnia, data = train_data, method = "class")

Matrice de confusion:

predicted_music_effect	Improve	No effect	Worsen
Improve	105	31	3
No effect	2	2	0
Worsen	0	0	0

Prédiction de l'effet de la musique sur la santé mentale

1. Arbre de décision

Métriques d'évaluations:

	'Improve' Class	'No effect' Class	'Worsen' Class
Accuracy (overall)	0.74		
Prévalence	0.74	0.23	0.02
Sensibilité	0.98	0.06	0
Spécificité	0.05	0.98	1
Balanced Accuracy	0.51	0.52	0.5

L'accuracy de ce modèle semble être d'environ 74%, mais cela peut être trompeur vu que le déséquilibre de classe n'est pas pris en compte.
Le taux de précision équilibré, offre une représentation plus précise de la performance du modèle

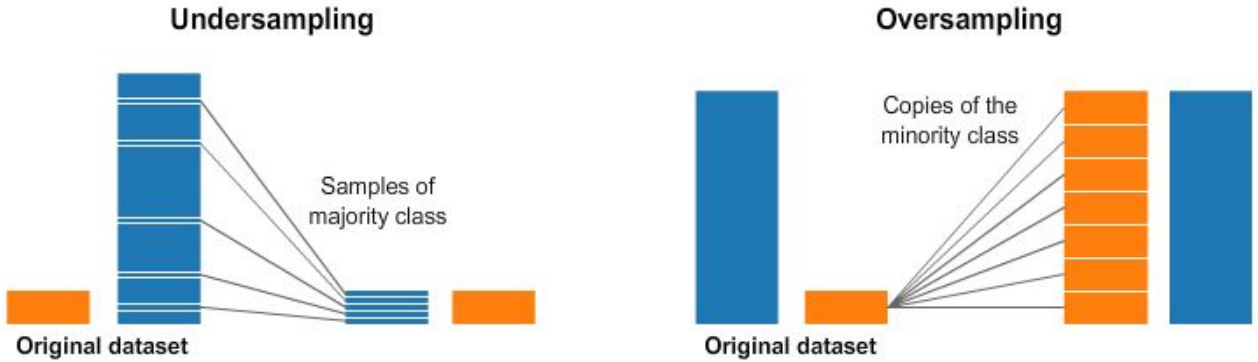
$$\text{Balanced Accuracy} = \frac{Sensitivity + Specificity}{2}$$

Where:

- Sensitivity (True Positive Rate) = $\frac{\text{True Positives}}{\text{True Positives}+\text{False Negatives}}$
- Specificity (True Negative Rate) = $\frac{\text{True Negatives}}{\text{True Negatives}+\text{False Positives}}$

Comment manipuler les jeu de données déséquilibrés?

1 - Méthodes d' oversampling & Undersampling



2 – Attribuer des poids plus élevés aux échantillons de classes minoritaires.

Modèle: Random Forest (avec des classes pondérées)
Bibliothèque utilisée: randomForest

```
rf_model <- randomForest(factor(Music.effects) ~ OCD + Anxiety + Depression + Insomnia,
                         data = train_data,
                         classwt = list(Improve = 1, `No effect` = 2, Worsen = 10))
```

Matrice de confusion:

predicted_music_effect	Improve	No effect	Worsen
Improve	94	28	2
No effect	13	5	1
Worsen	0	0	0

Métrique d'évaluation:

```
"Accuracy: 0.692307692307692"
"Balanced Accuracy: 0.510611205432937"
```

Prédiction de l'effet de la musique sur la santé mentale
3. Multinomial Logistic Regression

Modèle: Multinomial Logistic Regression

La régression logistique multinomial: prédire une variable catégorielle avec plusieurs niveaux.

Matrice de confusion:

	predictions		
	Improve	No effect	Worsen
Improve	204	17	0
No effect	51	12	0
Worsen	6	0	0

Métriques d'évaluation:

"Accuracy: 0.744827586206897"

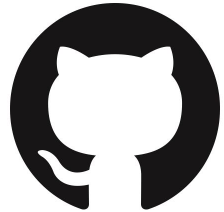
"Balanced Accuracy: 0.371184371184371"

Portfolio

- En global, il y’a des relations assez logiques dans notre étude.
- Certains facteurs (comme le BPM) ne jouent pas un rôle significatif.
- Jeu de données déséquilibré conduit à des conclusions biaisées du modèle.
- Pistes d'améliorations: plus d'instances pour une analyse plus robuste / ajout de questions sur les données démographiques telles que le pays, le sexe, etc.

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Portfolio



IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Conclusion

- En global, il y’a des relations assez logiques dans notre étude.
- Certains facteurs (comme le BPM) ne jouent pas un rôle significatif.
- Jeu de données déséquilibré conduit à des conclusions biaisées du modèle.
- Pistes d'améliorations: plus d'instances pour une analyse plus robuste / ajout de questions sur les données démographiques telles que le pays, le sexe, etc.

Merci pour votre attention

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom