**Case Description**

**Objective**

In the era of digital streaming, there’s an increasing need to categorize and recommend music based on genres. By analyzing various musical features extracted from tracks, we can delve deeper into their defining patterns. In this music genre classification project, you’ll work with a dataset containing various musical features extracted from tracks across different styles.

Please note that while this music genre classification dataset is extensive, it is incomplete. A significant portion of the records lacks specific genre information.

Your primary task is to predict the genres of these unlabeled tracks.

To accomplish this, you’ll employ Principal Component Analysis (PCA) to reduce the dimensionality of the dataset. By transforming the abundant features into principal components, you’ll streamline the data, making it more manageable and revealing patterns that are not immediately obvious in the raw data. The principal components you’ve derived will form the foundation for the next step in the project—employing a supervised machine learning algorithm, with a focus on the well-known logistic regression technique.

**Why Principal Component Analysis (PCA)?**

Music tracks are complex entities with numerous inherent features. Some of these features might be correlated. For instance, specific rhythm patterns might be prevalent in rock and blues. In this machine learning project, PCA can assist in reducing redundancy by transforming correlated musical features into a set of linearly uncorrelated variables or principal components. Reducing dimensionality can drastically improve the performance of classification algorithms by eliminating noise.

The features in the music genre classification dataset are designed to be intuitive and accessible allowing you to focus on the core concepts of PCA and machine learning without the need for specialized audio knowledge.

Prepare to dive deep into the layers of musical data and discover the patterns that help outline the musical genres.

**Project requirements**

You'll need Python 3 or a newer version and can choose any IDE (Jupyter Notebook, Spyder, PyCharm, Visual Studio, etc.). But be aware that the file provided with the solution for this music genre classification project is a Jupyter Notebook.

You’ll also need to have the following Python libraries installed:

* pandas
* NumPy
* matplotlib
* scipy
* scikit-learn
* seaborn

Pour votre projet de classification de genres musicaux, l’objectif est de prédire les genres de morceaux de musique en utilisant une analyse des caractéristiques musicales, la réduction de dimensions avec l’Analyse en Composantes Principales (PCA), et finalement un algorithme de machine learning supervisé, en particulier la régression logistique. Voici une approche détaillée pour structurer votre projet :

### Aperçu du Projet

1. **Exploration des Données :** Commencez par examiner le dataset pour comprendre la distribution de chaque caractéristique, détecter les valeurs manquantes, et identifier les schémas potentiels. Cette phase inclura également une analyse descriptive des variables.
2. **Analyse de Corrélation :** Une fois l'exploration de base effectuée, calculez les corrélations entre les variables. Cette étape est cruciale pour PCA, car elle aide à identifier les caractéristiques redondantes qui peuvent être réduites. Par exemple, il est probable que des variables comme Dynamic Range et Percussion Strength aient une certaine corrélation.
3. **PCA pour Réduction de Dimensionalité :**
   * Effectuez une normalisation des données, car PCA est sensible aux échelles des variables.
   * Appliquez PCA pour réduire le nombre de dimensions, tout en conservant la majorité de la variance. Cette réduction vise à simplifier le modèle tout en maintenant l’essentiel de l’information, en particulier en vue de rendre les prédictions plus performantes.
   * Visualisez les résultats de PCA en créant des graphiques, tels que des diagrammes de dispersion des deux premières composantes principales, pour voir comment les genres (lorsqu'ils sont étiquetés) se séparent dans cet espace réduit.
4. **Évaluation de l’Efficacité de la Classification – Données PCA vs Données d'Origine :**
   * Comparez les performances de la régression logistique (ou d’autres algorithmes de classification) en utilisant les données originales et les données transformées par PCA.
   * Utilisez des métriques telles que l’accuracy, la précision, le rappel et le F1-score pour mesurer les performances de votre modèle et déterminer l’effet de PCA sur la capacité de prédiction.
5. **Prédiction des Genres et Intégration :**
   * Une fois que vous avez un modèle performant, appliquez-le aux morceaux sans genre pour prédire leur catégorie.
   * Utilisez les prédictions pour générer des analyses, telles que des distributions de genres, et présentez ces résultats visuellement pour une interprétation plus intuitive.

### Aperçu des Caractéristiques de votre Dataset

Voici comment interpréter et exploiter les caractéristiques fournies :

* **Tempo (BPM)** : Influence souvent le genre (par ex., le rock a généralement un BPM plus élevé que l’ambient).
* **Dynamic Range** : Une large gamme dynamique peut être observée dans des genres comme le classique ou le jazz, où l’intensité varie fortement.
* **Vocal Presence** et **Percussion Strength** : Différentes genres mettent en avant certains éléments (ex. voix dans la pop, percussions dans l'électro).
* **Electronic Element Presence** et **Rhythm Complexity** : Des genres comme l’électro et le jazz se distinguent souvent par la présence électronique ou la complexité du rythme.
* **Drums Influence** et **Distorted Guitar** : Typiquement présents dans des genres tels que le rock ou le métal.
* **Instrumental Overlaps** et **Ambient Sound Influence** : Caractéristiques importantes dans les genres instrumentaux ou ambiants, pour ajouter de la texture et de la profondeur au son.

### Outils et Techniques

1. **Python Libraries :** Utilisez pandas pour la manipulation des données, matplotlib et seaborn pour la visualisation, sklearn.decomposition.PCA pour l’analyse en composantes principales, et sklearn.linear\_model.LogisticRegression pour le modèle de classification.
2. **Prétraitement :** Standardisez les caractéristiques numériques pour garantir que PCA et les modèles de machine learning fonctionnent efficacement.
3. **Evaluation** : Des méthodes comme la validation croisée vous permettront de mieux évaluer la généralisation du modèle.

Ce projet vous permettra de découvrir les liens subtils entre les caractéristiques musicales et les genres, en utilisant des techniques de machine learning pour explorer et interpréter des données musicales complexes.