Saint-Denis, Université Paris 8

Classification de genres musicaux par apprentissage automatique

Auteur :
DAUVERGNE Florian

Référent :
AMMI Mehdi

9 décembre 2024

Table des matières

Chapitre 1

Introduction

L'objectif de ce projet est de développer un modèle capable de classifier automatiquement des morceaux de musique en fonction de leur genre musical.

Cette tâche est complexe en raison de la diversité des genres et de la nature variée des données musicales.

Le projet explore deux algorithmes de machine learning populaires, **K-Nearest Neighbors (KNN)** et **Support Vector Classification (SVC)**, tout en intégrant des techniques d'optimisation comme le rééquilibrage des classes et la réduction dimensionnelle.

Les choix techniques sont guidés par l'efficacité des modèles, la gestion du déséquilibre des classes et la complexité des données, comme nous l'expliquerons dans les sections suivantes.

Chapitre 2

Préparation des Données

2.1 Analyse et Nettoyage des Données

Le nettoyage des données à été réalisé de la manière suivante :

— **Valeurs manquantes :** Les valeurs manquantes des variables numériques ont été remplacées par leur moyenne pour éviter de perdre des informations essentielles.

— **Colonnes non pertinentes :** Des colonnes comme le nom d'artiste et le titre de la chanson ont été supprimées car elles n'apportent pas de valeur pour la classification des genres.

Après le chargement du jeu de données, nous avons constaté une distribution inégale des genres, ce qui risque de biaiser les modèles en faveur des genres majoritaires.

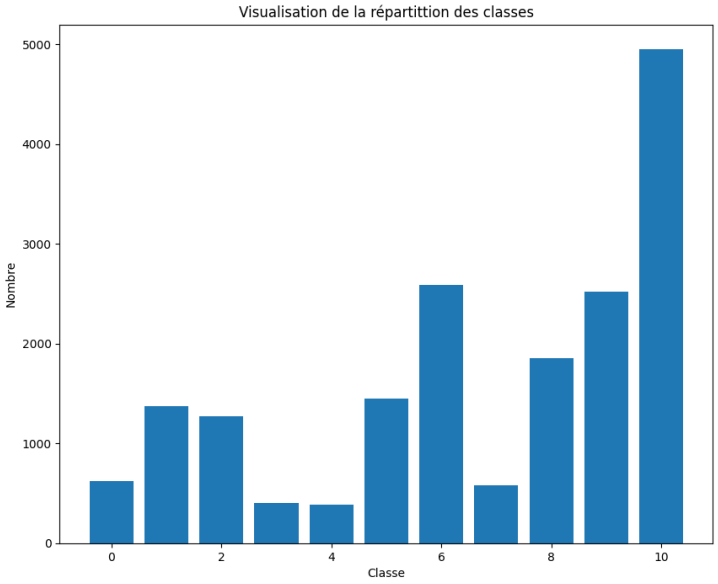

FIGURE 2.1 – Répartition des genres dans le jeu de données

2.2 Choix du Rééquilibrage des Données

Nous avons opté pour l'algorithme SMOTE (*Synthetic Minority Over-sampling Technique*) pour équilibrer les classes.

SMOTE génère des échantillons synthétiques pour les classes sous-représentées, ce qui aide les modèles à mieux apprendre les caractéristiques de chaque genre musical sans se concentrer uniquement sur les genres majoritaires.

Ce choix s'est avéré pertinent comme le montrent les résultats de précision obtenus après rééquilibrage (voir Section 5).

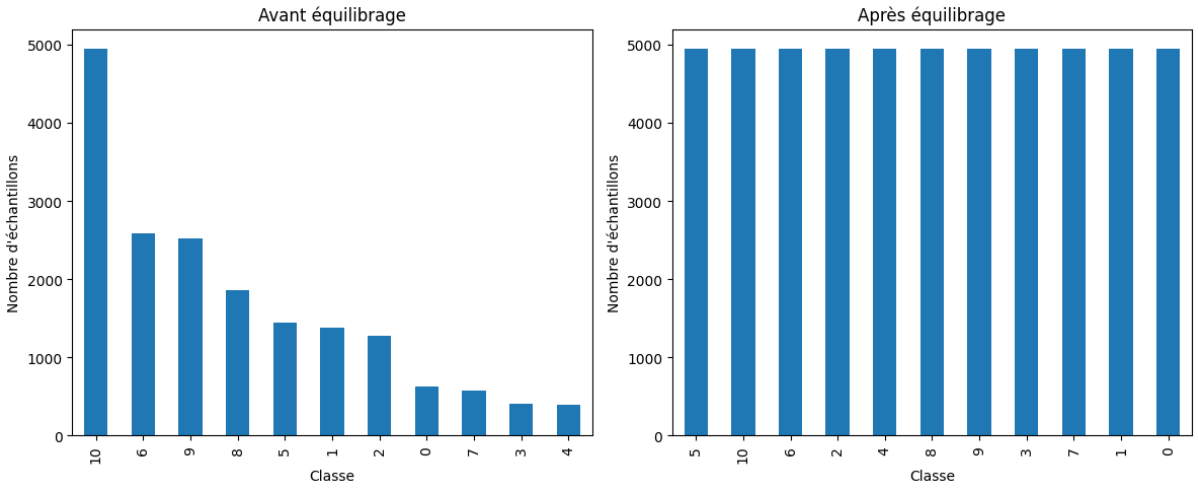

FIGURE 2.2 – Graphiques représentant la distribution avant et après rééquilibrage

Chapitre 3

Choix des Modèles et Optimisation des Hyperparamètres

3.1 Modèle K-Nearest Neighbors (KNN)

Le modèle KNN a été sélectionné pour sa simplicité et sa capacité à classifier en fonction de la proximité dans l'espace des caractéristiques.

Nous avons optimisé le nombre de voisins K en testant une plage de valeurs entre 1 et 30.

Les résultats montrent que la précision atteint un plateau à $K = 25$, ce qui a été retenu comme valeur optimale.

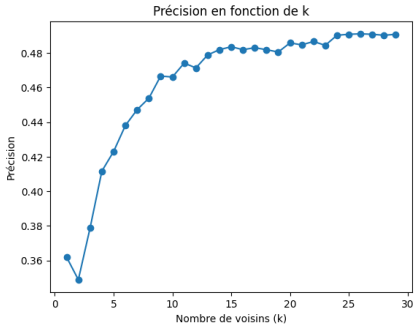

FIGURE 3.1 – Courbe de la précision du modèle KNN en fonction de K

3.2 Modèle Support Vector Classification (SVC)

Le modèle SVC a été choisi pour sa capacité à bien gérer des frontières complexes entre les classes, particulièrement utile pour des données musicales non linéaires.

Une recherche par grille a été effectuée pour ajuster les hyperparamètres de SVC (C, gamma et le type de noyau), en maximisant la précision globale.

Les valeurs optimales trouvées pour ces hyperparamètres ont permis à SVC de surpasser KNN dans les tests ultérieurs.

3.3 Résulats sans rééquilibrage

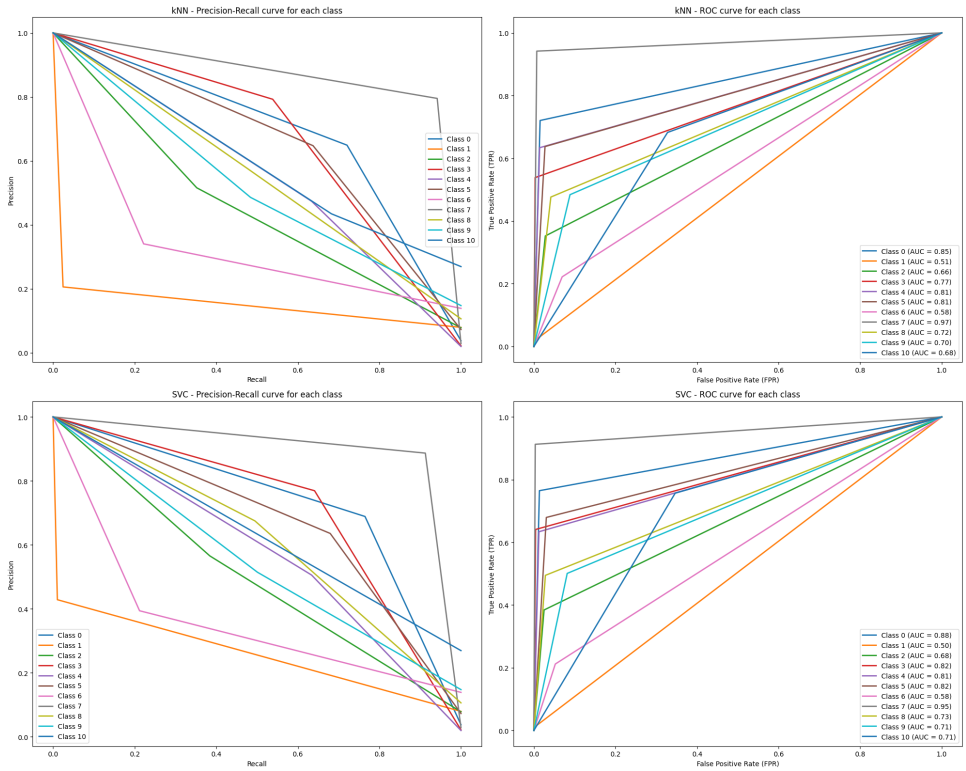

FIGURE 3.2 – Courbes ROC pour KNN et SVC avant rééquilibrage des données

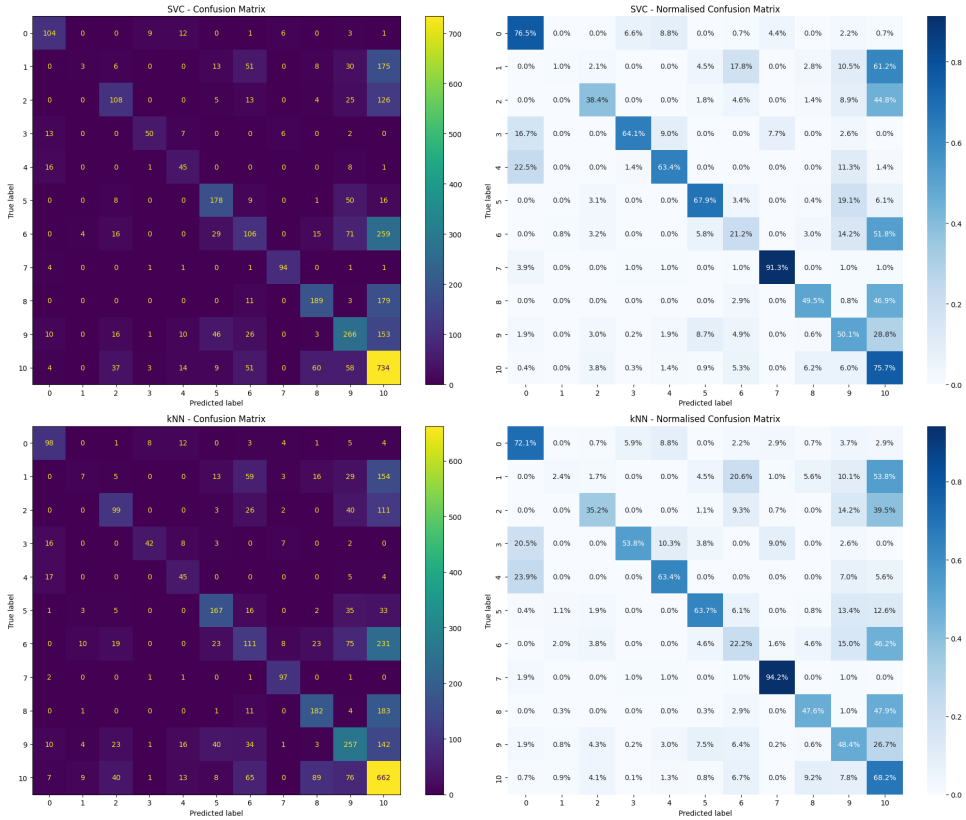

FIGURE 3.3 – Matrices de confusion avant rééquilibrage des données

Chapitre 4

Réduction Dimensionnelle par ACP

Pour réduire la complexité du modèle et minimiser le bruit dans les données, nous avons appliqué l'Analyse en Composantes Principales (ACP) après rééquilibrage des données.

ACP est particulièrement utile pour éviter le surapprentissage et accélérer le calcul en diminuant le nombre de dimensions.

Nous avons retenu les 9 premières composantes expliquant environ 85% de la variance des données, en veillant à maintenir une représentation significative des caractéristiques.

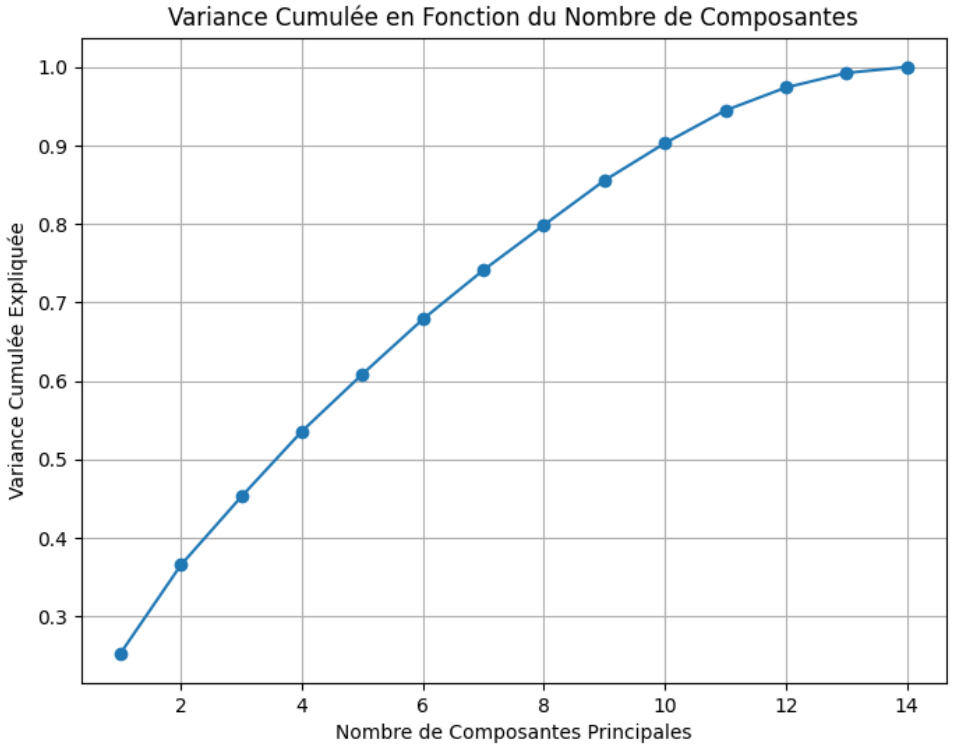

FIGURE 4.1 – Variance cumulée en fonction du nombre de composantes

Bien que la réduction de dimensions ait légèrement réduit la précision, elle a amélioré la généralisation et réduit la complexité sans affecter considérablement les performances.

Chapitre 5

Évaluation et Résultats

5.1 Performance sur les Données d'Origine

Les premières évaluations sur les données d'origine montrent que le modèle SVC surpasse KNN, surtout dans la classification des classes majoritaires.

Cependant, les résultats sur les classes minoritaires sont faibles en raison du déséquilibre initial.

5.2 Impact du Rééquilibrage avec SMOTE

Après rééquilibrage, le modèle SVC améliore sa précision pour les classes sous-représentées, tandis que KNN bénéficie également d'un léger gain en performance.

La Figure 5.1 montre les courbes ROC et les scores AUC pour chaque classe, soulignant la meilleure capacité de SVC à discriminer les genres après équilibrage.

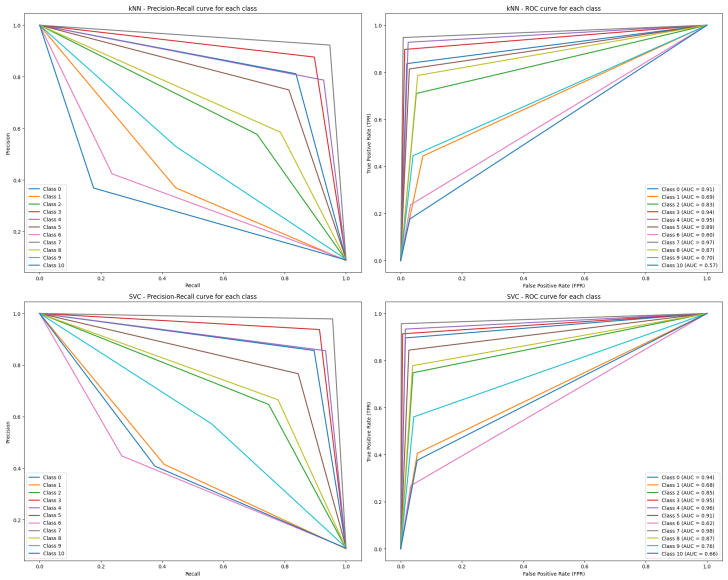

FIGURE 5.1 – Courbes ROC pour KNN et SVC après rééquilibrage des données

5.3 Réduction Dimensionnelle et Évaluation Finale

La réduction des dimensions via ACP a permis d'améliorer la vitesse de calcul et de réduire le risque de surapprentissage sans trop diminuer la précision du modèle.

Les résultats finaux montrent que SVC conserve une précision élevée avec seulement 9 dimensions principales.

5.4 Comparaison Finale des Modèles

Les résultats finaux révèlent que le modèle SVC est supérieur à KNN en termes de précision et de rappel, particulièrement sur les classes rééquilibrées.

TABLE 5.1 – Résultats finaux des modèles SVC et KNN sur les données équilibrées et réduites

Modèle	Précision	Rappel	F1-Score
SVC	0.69	0.70	0.69
KNN	0.64	0.66	0.64

Chapitre 6

Conclusion et Recommandations

Ce projet a mis en lumière l'importance des choix techniques dans la classification des genres musicaux :

- **Modèle préféré** : Le modèle SVC s'est avéré supérieur à KNN pour ce projet, en particulier après équilibrage des classes et réduction des dimensions.
- **Rééquilibrage des données** : Le rééquilibrage via SMOTE a significativement amélioré la précision pour les classes sous-représentées.
- **Réduction des dimensions** : La réduction dimensionnelle par ACP a permis d'améliorer l'efficacité sans compromettre les performances.

Pour des développements futurs, il est recommandé de tester d'autres modèles, comme les forêts aléatoires ou les réseaux de neurones et d'explorer des techniques avancées d'optimisation et de traitement des données pour de meilleurs résultats.