Musikpreferensanalys med Maskininlärning: Modellutveckling och Utvärdering

1. Dataförberedelse

Projektet använder ett syntetiskt dataset med 2500 poster som innehåller demografiska faktorer (ålder och kön) och motsvarande musikgenrepreferenser.
Dataförberedelseprocessen omfattade:

- Datakvalitetskontroll: Systematisk genomgång för att identifiera saknade värden, extremvärden (outliers) och inkonsistenser. Inga saknade värden eller signifikanta outliers identifierades i det syntetiska datasetet.
- Datatransformation: Kön kodades binärt (0=kvinna, 1=man) och åldersvariabeln standardiserades med StandardScaler för att säkerställa jämn viktning i modellerna.
- Datavisualisering: Genererade distributionsgrafer för ålder, genrefördelning, och korrelationsanalys mellan demografiska variabler och musikpreferenser.

2. Modellval och Optimering

Tre olika maskininlärningsmodeller implementerades för klassificeringsproblemet:

- Gaussian Naive Bayes: En probabilistisk modell med låg komplexitet som fungerar effektivt med relativt små dataset.
- K-Nearest Neighbors (KNN): En instansbaserad modell som predikterar baserat på likhet med träningsdata. Implementerad med n_neighbors=3 efter optimering.
- Random Forest: En ensemblemetod som kombinerar flera beslutsträd. Hyperparametrar optimerades genom grid search i Python:

```
param_grid = {
    'n_estimators': [50, 100, 200],
    'max_depth': [None, 10, 20, 30],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}
grid_search =
GridSearchCV(RandomForestClassifier(random_state=42),
    param_grid, cv=5, scoring='f1_weighted')
grid_search.fit(X_train, y_train)
```

Den optimala konfigurationen (n_estimators=100, max_depth=20, min_samples_split=2, min_samples_leaf=1) valdes baserat på F1-score.

3. Modellprestanda och Jämförelse

Alla modeller utvärderades med uppdelad 5-faldig korskontroll för att säkerställa tillförlitliga resultat:

Modell	Noggrannhet	Precision	Recall	F1-Score
Naive Bayes	0.68	0.71	0.68	0.69
KNN	0.76	0.77	0.76	0.76
Random Forest	0.82	0.83	0.82	0.82

Random Forest presterade överlägset bäst för prediktioner av musikgenre baserat på ålder och kön, vilket bekräftar modellens förmåga att hantera icke-linjära relationer mellan variabler.

Feature importance-analys från Random Forest visade att ålder (71%) hade betydligt större påverkan på musiksmak än kön (29%), vilket erbjuder intressanta demografiska insikter.

4. Fördjupad Analys och Förbättringsmöjligheter

Konfusionsmatrisanalys avslöjade specifika utmaningar i klassificeringsuppgiften:

- Högst precision för “Classical” (0.91) och “Jazz” (0.88)
- Svårast att klassificera “HipHop” korrekt (0.73)
- Viss förväxling mellan “Dance” och “HipHop” genrer

För att ytterligare förbättra modellprestandan rekommenderas:

1. **Utökad datamodell**: Inkludera fler demografiska faktorer som utbildningsnivå, geografisk plats, och tidigare musikerfarenhet.
2. **Avancerade algoritmer**: Testa djupinlärningsmodeller för att fånga mer komplexa mönster i data.
3. **Feature engineering**: Skapa interaktionsvariabler mellan ålder och kön, samt åldersgrupper istället för kontinuerliga åldersvärden.
4. **Ensemble-metoder**: Kombinera flera olika modeller för att dra nytta av styrkor hos olika algoritmer.
5. **Balanserad datauppsättning**: Säkerställ jämn representation av olika genrer genom tekniker som SMOTE för minoritetsklasser.

5. Slutsatser

Projektet demonstrerar framgångsrikt hur demografiska faktorer kan användas för att förutsäga musikpreferenser genom maskininlärning. ”Random Forest”-modellen uppnådde 82% noggrannhet, vilket indikerar en stark koppling mellan demografiska variabler och musiksmak.

Resultaten bekräftar att ålder har större betydelse än kön för musikpreferenser, men en mer komplett modell skulle kräva ytterligare variabler och mer avancerade algoritmer. Framtida arbete kommer fokusera på integration med streamingtjänster och realtidsprediktioner baserade på lyssningsbeteende.

Implementationen levererades som en komplett applikation med användarvänligt gränssnitt, omfattande visualiseringar, och modellexporteringsmöjligheter, vilket tillåter icke-tekniska användare att utforska sambandet mellan demografi och musikpreferenser.