Musikpreferensanalys med Maskininlärning - Projektöversikt

1. Problemställning

Projektet syftar till att utveckla en maskininlärningsmodell som kan förutsäga musikgenrepreferenser baserat på demografiska faktorer, primärt ålder och kön. Målet är att skapa ett verktyg som kan:

- Analysera sambandet mellan demografiska egenskaper och musikpreferenser
- Förutsäga vilken musikgenre en person sannolikt föredrar baserat på ålder och kön
- Visualisera trender i hur musiksmak varierar över åldersgrupper och mellan könen

2. Data

För projektet används ett syntetiskt dataset bestående av 2500 datapunkter med följande struktur:

- **Ålder**: Numerisk variabel (integer)
- **Kön**: Binär kodning (0 = kvinna, 1 = man)
- **Genre**: Kategorisk variabel för musikgenrepreferens

2.1 Datakvalitet

- **Komplett data**: Det syntetiska datasetet innehåller inga saknade värden, vilket säkerställer fullständiga träningsdata.
- **Null-värden**: Applikationen implementerar metoder för att upptäcka och hantera null-värden med olika imputeringsstrategier, trots att det syntetiska datasetet inte har några.
- **Extrema värden**: Programmet identifierar åldersoutliers med IQR-metoden (värden utanför Q1-1,5×IQR eller Q3+1,5×IQR).
- **Datatyper**: Datasetet innehåller två feature-datatyper: ålder (kontinuerlig numerisk) och kön (binär kategorisk). Målvariabeln ‘genre’ är kategorisk.

2.2 Dataförbehandling

- Ålder standardiseras med StandardScaler för att säkerställa lika viktning i avståndsbaserade algoritmer.
- Kön används som en binär feature (0=kvinna, 1=man).
- Musikgenrer behålls som kategoriska etiketter för klassificering.
- Stratifierad tränings-/testuppdelning med justerbar proportion (standardvärde 30%).

3. Problemtyp och Metodik

Detta är ett **klassificeringsproblem** där modellen ska prediktera en kategorisk variabel (musikgenre) baserat på demografiska egenskaper. Projektet använder supervised learning med labeled data.

Tre olika modeller implementeras för att jämföra prestanda:

1. **Gaussian Naive Bayes**: En probabilistisk klassificerare som är effektiv för mindre dataset och fungerar väl med få samples per klass.
2. **K-Nearest Neighbors (KNN)**: En instansbaserad metod som klassificerar baserat på likhet med träningsexempel, med n_neighbors=3.
3. **Random Forest**: En ensemblemetod som kombinerar flera beslutsträd, robust mot outliers och överanpassning, implementerad med n_estimators=100.

4. Implementation

- **Arkitektur**: Objektorienterad design med MusicPreferencesApp-klassen som central komponent.
- **Teknologi**: Python med scikit-learn, pandas, numpy, matplotlib, seaborn och tkinter.
- **Användargränssnitt**: Tab-baserat GUI med flikar för datautforskning, modellträning, prediktion och information.
- **Datautforskning**: Visualiseringar inkluderar åldersfördelning, genrefördelning, ålder per genre, och genrepreferenser uppdelat efter kön.
- **Utvärderingsmetriker**: Precision, recall, F1-score och konfusionsmatris, samt feature importance för Random Forest.

5. Resultat och Slutsatser

Projektet demonstrerar ett fullständigt maskininlärningsarbetsflöde från datainsamling och analys till modellträning och implementering i en användarvänlig applikation. Även med begränsad demografisk data kan maskininlärningsmodeller göra rimliga prediktioner av musikpreferenser, med Random Forest som generellt presterar bäst tack vare dess förmåga att fånga icke-linjära relationer.

Framtida utvecklingsmöjligheter inkluderar utökning av datamodellen med fler demografiska faktorer, implementation av avancerade modeller som deep learning, och integration med musikstreamingtjänster för realtidsprediktioner baserat på lyssningsbeteende.