Laporan Mata Kuliah Pembelajaran Mesin

Music & Mental Health Survey Results

disusun untuk memenuhi

tugas Pembelajaran Mesin

oleh :

Kelompok 4

Glenn Hakim	2208107010072
Ahmad Syah Ramadhan	2208107010033
Andika Pebriansyah	2208107010058
Nisa Rianti	2208107010018
Nuri Masyithah	2208107010006

JURUSAN INFORMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS SYIAH KUALA

2025

BAB I
PENDAHULUAN

1.1 Latar Belakang

Musik memiliki peran penting dalam kesejahteraan emosional dan mental. Terapi musik (Music Therapy/MT) digunakan untuk meredakan stres dan meningkatkan suasana hati melalui stimulasi hormon seperti oksitosin. Namun, efektivitas MT dapat bervariasi tergantung pada genre musik dan preferensi individu.

Kumpulan data Music & Mental Health (MxMH) bertujuan untuk mengeksplorasi hubungan antara selera musik seseorang dan kondisi kesehatan mental yang mereka laporkan. Temuan ini diharapkan dapat mendukung penerapan MT yang lebih efektif dan memberikan wawasan mengenai dampak musik terhadap kesehatan mental.

Tugas ini dirancang untuk memberikan pengalaman praktis dalam mengolah data dari sumber open source, seperti Kaggle dan Hugging Face. Melalui berbagai tahap data preparation, diharapkan dapat memahami pentingnya proses ini serta menerapkan teknik yang tepat untuk meningkatkan kualitas data yang digunakan dalam analisis atau pengembangan model machine learning.

1.2 Tujuan

1. Memahami cara memilih dan memuat dataset dari sumber open source.
2. Melihat relasi antara lama waktu mendengarkan musik dan jenis musik tertentu terhadap perubahan kesehatan mental.
3. Menganalisis struktur serta karakteristik dataset yang digunakan.
4. Mengimplementasikan teknik preprocessing untuk meningkatkan kualitas dataset, termasuk penanganan missing values, encoding, normalisasi, dan feature selection.

BAB II
PEMBAHASAN

2.1 Deskripsi Dataset

Dataset *Music x Mental Health Survey Results* dari Kaggle mengeksplorasi hubungan antara preferensi musik dan kesehatan mental, mencakup 736 responden dengan 33 fitur.

- Blok 0: Informasi Umum – Kebiasaan mendengarkan musik dan faktor yang berpotensi mempengaruhi kesehatan mental.
- Blok 1: Preferensi Musik – Frekuensi mendengarkan 16 genre musik, seperti Pop, Rock, Hip-hop, dan Classical, dalam skala *Never* hingga *Very frequently*.
- Blok 2: Kesehatan Mental – Responden menilai tingkat anxiety, depression, insomnia, dan OCD pada skala 0-10.
- Fitur Tambahan – Usia, genre favorit, durasi mendengarkan musik, serta platform streaming yang digunakan.
- Fitur Target – *Music effects* menunjukkan dampak musik terhadap kesehatan mental dengan kategori *Improve, No effect,* atau *Worsen*.

Dataset ini dikumpulkan oleh @catherinerasgaitis melalui Google Form dan disebarkan di Reddit, Discord, media sosial, serta tempat umum. Survei dirancang agar ringkas dan mudah diisi untuk meningkatkan jumlah responden.

2.2 Data Loading

Proses memuat dataset ini dilakukan dengan memanfaatkan library kagglehub untuk mendownload dataset ini langsung dari platform kaggle dan membacanya dengan bantuan library pandas dari python, berikut adalah implementasi kodenya

```
path = kagglehub.dataset_download("catherinerasgaitis/mxmh-survey-results")

data = pd.read_csv(path + "/mxmh_survey_results.csv")
data.head(10)
```

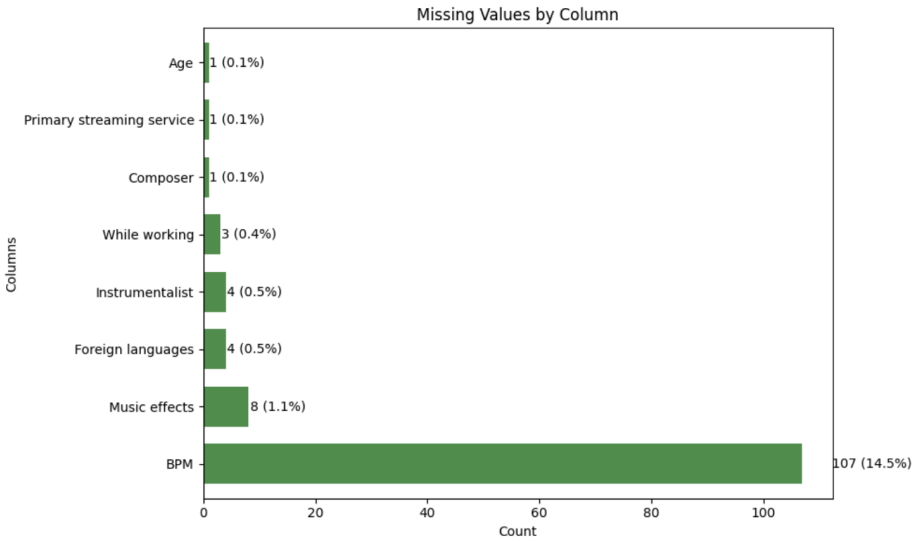
Gambar 1. *Kode python untuk memuat dataset dengan kagglehub dan pandas*

Pada proses ini, pemuatan data berjalan dengan lancar tanpa kendala maupun tantangan yang berarti. Semua tahapan, mulai dari mengunduh dataset menggunakan library kagglehub hingga membacanya dengan pandas, dapat dilakukan secara optimal tanpa hambatan teknis. Ini menunjukkan bahwa lingkungan pemrograman telah dikonfigurasi dengan baik, koneksi internet stabil, serta dependensi yang diperlukan telah terpasang dengan benar.

2.3 Data Understanding

Tahap pemahaman data dilakukan dengan berbagai cara untuk mendapatkan informasi dari eksplorasi awal data. Dengan memanfaatkan metode-metode statistik dasar dan visualisasi diagram, berikut adalah proses pemahaman data yang kami lakukan

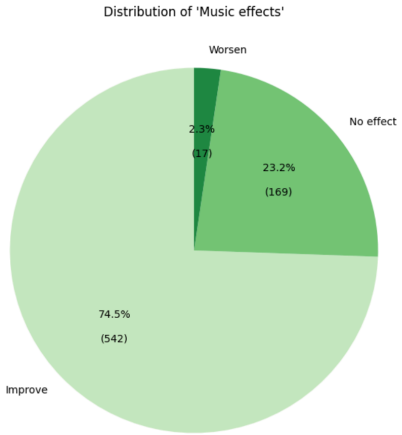
a. Missing Values



Gambar 2. *Bar plot missing values*

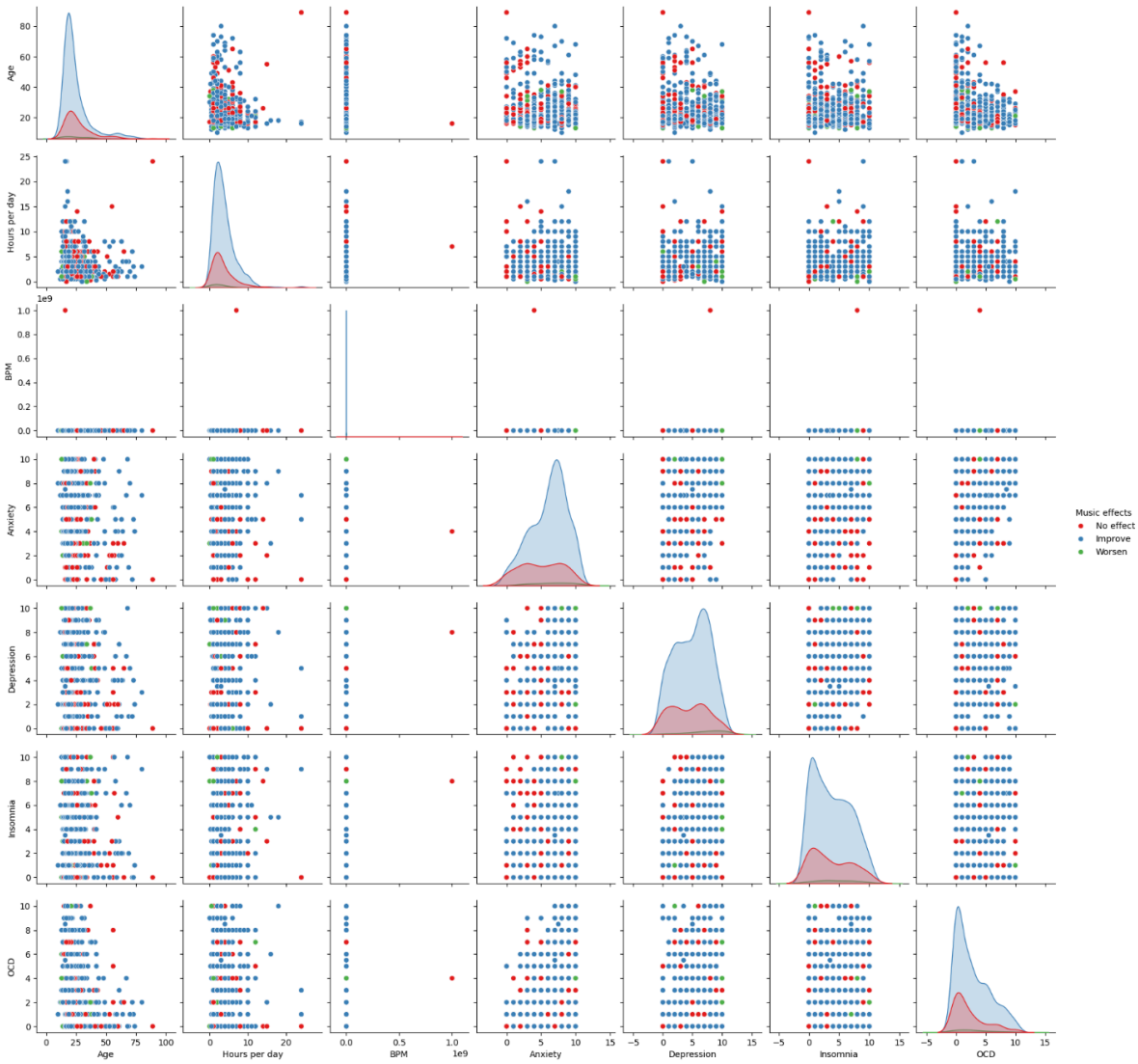
Dari bar plot diatas dapat dilihat bahwa ada nilai yang hilang pada beberapa kolom di dataset. Diantara kolom tersebut yang paling banyak mempunyai *missing values* adalah kolom BPM (Beat Per Minute), yaitu sebanyak 107 baris data yang hilang atau sama dengan 14.5% dari jumlah keseluruhan data.

b. Distribusi Data



Gambar 3. *Pie chart Distribusi kelas target (Music effects)*

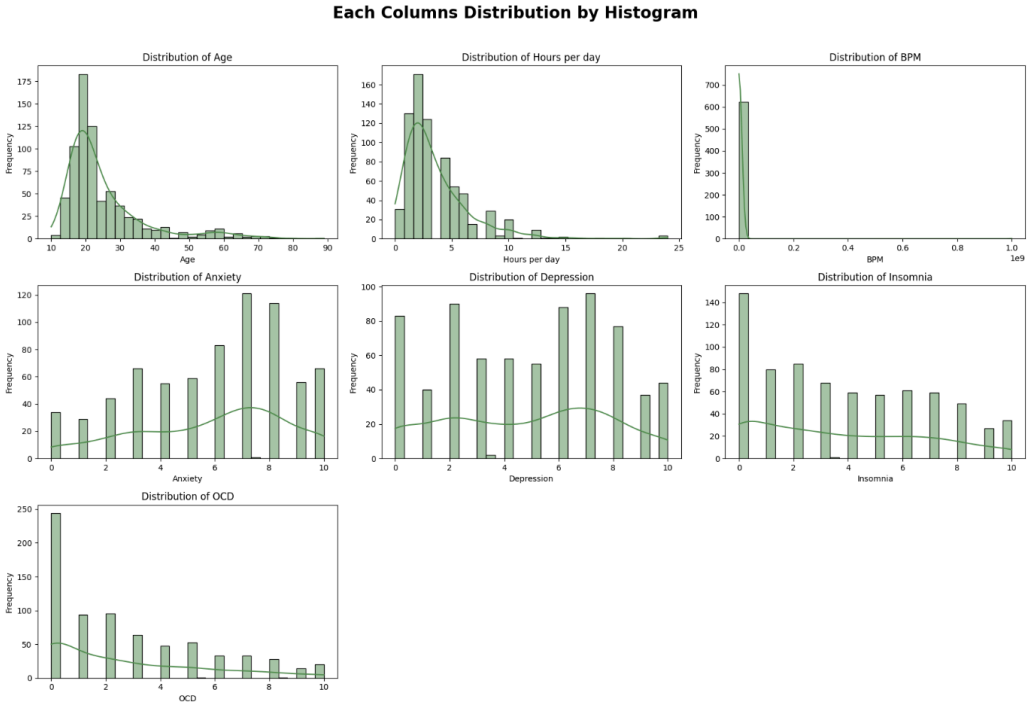
Diagram diatas menunjukkan distribusi fitur target pada dataset ini yaitu *Music effects*. Terdapat 3 kategori kelas pada fitur target dan terlihat sangat jelas bahwa persebaran kelas tidak seimbang, sangat condong ke kategori kelas *Improve* yaitu sebanyak 74.5% dari persentase keseluruhan data. Kemudian diikuti dengan kategori kelas *No effects* yaitu dengan persentase 23.2% dari keseluruhan data. Serta kategori kelas *Worsen* dengan persentase paling sedikit yaitu 2.3% dari jumlah keseluruhan data.



Gambar 4. *Scatter Plot distribusi data tiap pasangan kolom berdasarkan Music effects*

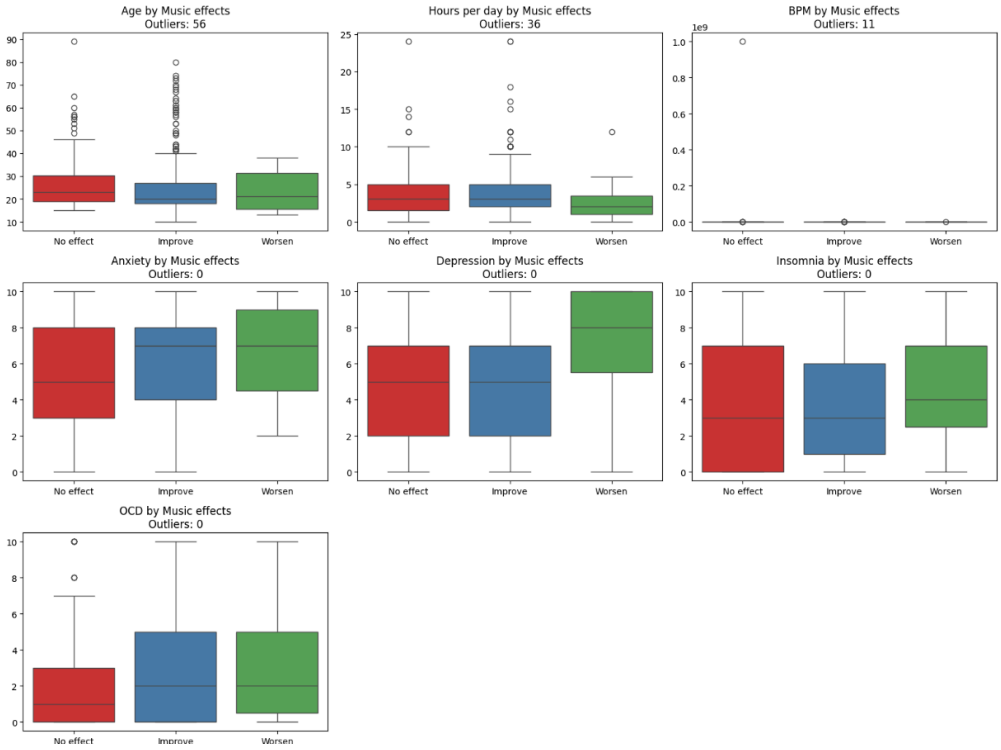
Scatter Plot menunjukkan bagaimana distribusi data setiap pasangan kolom dengan pengelompokan warna pada fitur target. Fungsi dari scatter plot ini adalah untuk mengidentifikasi pola, hubungan, atau tren antara dua variabel, serta melihat sejauh mana fitur-fitur dapat membedakan kategori pada target. Dengan pengelompokan warna berdasarkan fitur target, scatter plot juga membantu dalam mengevaluasi separabilitas kelas, yang berguna dalam pemilihan fitur dan pemahaman karakteristik dataset sebelum

diterapkan ke model machine learning.



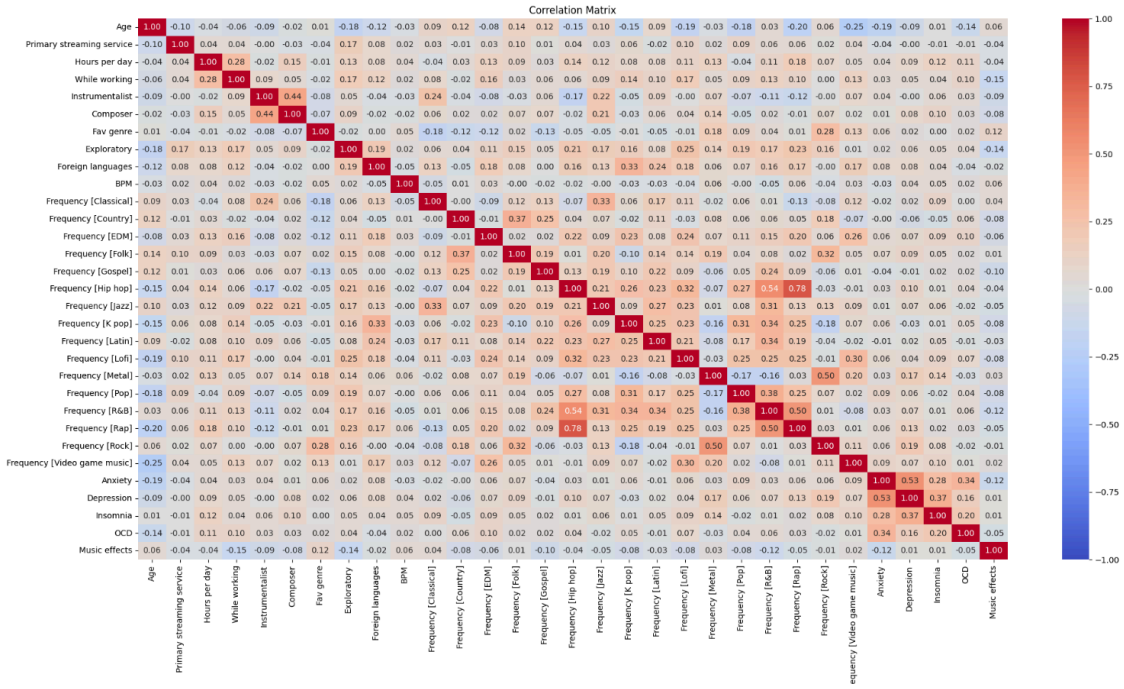
Gambar 5. *Histogram distribusi tiap kolom numerical dengan*

Pada visualisasi diatas dapat terlihat bahwa fitur BPM mempunyai distribusi data yang sangat miring dibandingkan fitur-fitur lainnya, ini menunjukkan bahwa persebaran data pada kolom BPM kurang baik dan harus lebih banyak dilakukan pra-pemrosesan lebih lanjut, seperti melakukan *outliers handling,* normalisasi, sampai dengan *skewness handling* seperti transformasi log jika diperlukan.



Gambar 6. *Box plot nilai outliers pada fitur numeric dengan pengelompokan pada fitur target (Music effects) menggunakan*

Pada visualisasi di atas, terlihat bahwa beberapa fitur memiliki distribusi data yang cukup baik, sementara yang lain menunjukkan adanya skewness dan outlier yang perlu ditangani. Age dan Hours per Day memiliki banyak outlier, sedangkan BPM menunjukkan distribusi sangat miring dengan outlier ekstrim, memerlukan normalisasi atau transformasi. Sementara itu, fitur Anxiety, Depression, Insomnia, dan OCD lebih seimbang, meski OCD memiliki beberapa outlier. Menariknya, kategori "Worsen" pada **Depression** memiliki median lebih tinggi, mengindikasikan musik dapat memperburuk kondisi depresi bagi sebagian orang. Oleh karena itu, fitur dengan skewness tinggi dan banyak outlier memerlukan pra-pemrosesan lebih lanjut agar analisis lebih akurat.



Gambar 7. *Correlation matrix heatmap antar tiap kolom setelah encoding*

Matriks korelasi ini menunjukkan hubungan antar fitur dalam dataset, dengan warna merah menunjukkan korelasi positif dan biru menunjukkan korelasi negatif. Usia memiliki korelasi negatif dengan frekuensi mendengarkan musik video game (-0.25) dan lofi (-0.19), menunjukkan bahwa individu yang lebih tua cenderung jarang menikmati genre tersebut. Sebaliknya, jumlah jam mendengarkan musik per hari berkorelasi positif dengan hip-hop (0.31), jazz (0.22), dan EDM (0.16), yang menunjukkan preferensi lebih tinggi terhadap genre-genre ini.

Beberapa hubungan yang cukup kuat terlihat antara hip-hop dan R&B (0.53), serta antara pop dengan rock (0.38) dan R&B (0.34). Sementara itu, fitur terkait kesehatan mental seperti anxiety, insomnia, dan depression tidak menunjukkan korelasi yang signifikan dengan kebiasaan mendengarkan musik, tetapi OCD memiliki korelasi negatif dengan usia (-0.14). Secara keseluruhan, meskipun tidak ada korelasi yang sangat kuat, terdapat pola menarik dalam preferensi musik dan kebiasaan mendengarkan berdasarkan fitur yang dianalisis.

2.4 Data Preparation

Persiapan data merupakan tahap krusial dalam data mining yang menentukan kualitas model. Berdasarkan pemahaman data sebelumnya, kami melakukan berbagai proses pra-pemrosesan untuk memastikan data yang bersih dan siap digunakan dalam pembangunan model.

a. Missing Values Handling

Penanganan nilai hilang dilakukan sesuai dengan tipe data, distribusi, dan *domain knowledge* masing-masing fitur.

- Fitur Age

```
mean_hiphop_age = round(data_preprocessed[data_preprocessed['Fav genre'] == "Hip hop"]['Age'].mean())

print("Mean Hip Hop Age: ", mean_hiphop_age)

data_preprocessed['Age'] = data_preprocessed['Age'].fillna(value = mean_hiphop_age)
data_preprocessed['Age'].isnull().sum()
```

Gambar 8. *Implementasi kode untuk mengatasi nilai hilang pada fitur Age*

Karena nilai yang hilang terdapat pada pengguna dengan genre favorit Hip Hop, usia yang hilang diisi menggunakan rata-rata usia pendengar Hip Hop. Pendekatan ini lebih relevan secara kontekstual dibandingkan imputasi global.

- Fitur Foreign languages, While working, Instrumentalist, dan Primary streaming service

```
column_to_fill = ['Foreign languages', 'While working', 'Instrumentalist', 'Primary streaming service']
for column in column_to_fill:
    data_preprocessed[column] = data_preprocessed[column].fillna(data_preprocessed[column].mode().values[0])

print("Success filling missing values")
```

Gambar 9. *Implementasi kode mengatasi nilai hilang pada fitur* yang dipilih

Nilai yang hilang pada fitur Foreign languages, While working, Instrumentalist, dan Primary streaming service diisi menggunakan modus (mode), yaitu nilai yang paling sering muncul. Metode ini dipilih karena fitur-fitur tersebut bersifat kategorikal, sehingga lebih relevan dibandingkan menggunakan mean atau median yang lebih cocok untuk data numerik.

- Fitur BPM

```
mean_bpm_dict = data_preprocessed.groupby('Fav genre')['BPM'].mean().round().to_dict()

for genre in data_preprocessed['Fav genre'].unique():
    data_preprocessed.loc[data_preprocessed['Fav genre'] == genre, 'BPM'] = data_preprocessed[data_preprocessed['Fav genre'] == genre]
    ['BPM'].fillna(value = mean_bpm_dict[genre])

print("Success filling missing values")
```

Gambar 10. *Implementasi kode mengisi nilai hilang BPM*

Fitur BPM menggambarkan jumlah ketukan per menit dari genre musik favorit seseorang. Hasil analisis menunjukkan distribusi yang sangat miring (skewed), dengan nilai ekstrim seperti 0, 624, dan 999999999 yang tidak masuk akal dalam konteks musik dan dianggap sebagai outlier. Secara domain knowledge, BPM musik umumnya berada dalam rentang tertentu, sehingga nilai-nilai tersebut diubah menjadi NaN agar tidak memengaruhi analisis. Selanjutnya, missing values diisi menggunakan rata-rata BPM berdasarkan genre musik favorit masing-masing pengguna, sehingga lebih akurat dibandingkan menggunakan rata-rata seluruh dataset.

- Fitur Target (Music effects)

```
data = data.dropna(subset=['Music effects'])
print("Missing Values on Music effects: ", data['Music effects'].isnull().sum())
```

Gambar 11. *Implementasi kode menghapus missing values fitur Music effects*

Fitur Music effects merupakan target variabel dalam dataset ini, sehingga data dengan nilai yang hilang pada fitur tersebut dihapus. Langkah ini diambil karena fitur target harus memiliki nilai yang lengkap agar model dapat belajar dengan baik, serta tidak dapat diimputasi tanpa menimbulkan bias dalam prediksi. Setelah penghapusan, dilakukan pengecekan ulang untuk memastikan bahwa tidak ada lagi nilai yang hilang pada fitur ini.

b. Encoding Categorical Features

```
import numpy as np
from sklearn.preprocessing import LabelEncoder

data_preprocessed = data_preprocessed.replace("", np.nan)

label_encoders = {}
for column in categorical_features:
    le = LabelEncoder()

    non_null_data = data_preprocessed[column].dropna()
    le.fit(non_null_data)

    data_preprocessed.loc[data_preprocessed[column].notna(), column] = le.transform(non_null_data)

    label_encoders[column] = le

data_preprocessed.head()

```

Gambar 12. *Implementasi kode untuk encoding*

Untuk mengonversi data kategorik menjadi numerik, kami menggunakan Label Encoding dari sklearn.preprocessing.

- Langkah pertama, nilai kosong ("") dalam dataset diubah menjadi NaN.
- Setiap fitur kategorikal dikodekan menggunakan LabelEncoder, yang mengubah kategori menjadi angka.
- Proses encoding hanya diterapkan pada data non-null agar tidak mengubah nilai NaN.
- Hasil encoding disimpan dalam dictionary (label_encoders) untuk referensi jika diperlukan kembali.

Metode ini digunakan karena Label Encoding sederhana dan cocok untuk fitur dengan nilai kategori ordinal atau tanpa hubungan kompleks antar kategori.

c. Feature Selection atau Extraction

```
1  data = data.drop(columns=['Timestamp', 'Permissions'])
```

Gambar 13. *Implementasi kode untuk menghapus fitur Timestamp dan Permissions*

Dalam proses ini, kami menghapus dua fitur, yaitu Timestamp dan Permissions, dengan alasan sebagai berikut:

- Permissions memiliki variance 1, yang berarti semua nilai dalam kolom ini identik. Karena tidak ada variasi, fitur ini tidak memberikan informasi yang berguna untuk analisis atau prediksi.
- Timestamp hanya berisi informasi tentang waktu pengisian formulir, yang merupakan metadata dan tidak berkontribusi terhadap pemahaman hubungan antara musik dan kesehatan mental.

Penghapusan kedua fitur ini bertujuan untuk mengurangi dimensi data, menghilangkan informasi yang tidak relevan, serta meningkatkan efisiensi model dalam proses analisis dan pelatihan.