**Spotify Song Prediction**

**The Problem**

Saat ini kemampuan untuk memprediksi sesuatu yang mungkin akan menjadi popular adalah riset yang penting bagi setiap industri. Terutama sekali bagi pertumbuhan dan kompetisi pada industri music. Sejak meluasnya penggunaan platform musik digital (Spotify, Billboard, Lastfm), data sangatlah mudah dijangkau dan perilaku pendengar bisa dengan mudah diobservasi. Hal ini memberikan kemudahan bagi teknik peramalan dan juga kerapkali digunakan pada sistem rekomendasi.

**Goals & Success**

Metrik yang digunakan adalah popularity. The Spotify Popularity Index memiliki range dari 0-sampai-100 yang menunjukkan peringkat seberapa popular seorang artis atau lagu relative terhadap artis dan track pada spotify.

**Key Solution**

Untuk melakukan analisa prediktif, didapatkan dataset dari

Dataset yang digunakan adalah sebagai berikut :

|  |  |
| --- | --- |
| 1. Accousticness : | ukuran dari 0.0 to 1.0 apakah suatu track adalah akusting, dengan 1.0 merepresentasikan angka keyakinan yang tinggi bahwa suatu track adalah akustik. |
| 1. Danceability : | Mendeskripsikan seberapa cocok suatu track untuk digunakan berdasa berdasarkan kombinasi dari elemen music termasuk tempo, stabilitas ritme, kekuatan beat dan keseluruhan Nilai 0.0 adalah kurang cocok untuk berdansa dan 1.0 adalah sangat cocok |
| 1. Duration : | Panjang track dalam millisecond |
| 1. Energy : | Ukuran dari 0.0 sampai dengan 1.0 yang merepresentasikan ukuran intensitas dan aktifitas. Track yang berenergi akan terasa capat dan riuh. |
| 1. Explicit | Apakah suatu track memiliki lirik yang eksplisit |
| 1. ID : | ID Spotify untuk suatu track |
| 1. Instumentalness : | Memprediksi agar suatu track tidak memiliki |
| 1. Key : | Kunci yang digunakan suatu track, menggunakan notasi standard pitch |
| 1. Liveliness : | Mendeteksi keberadaan penonton di dalam rekaman. Nilai yang tinggi menunjukkan kemungkinan bahwa track direkam secara langsung. |
| 1. Loudness : | Tingkat kebisingan track dalam decibel (db) |
| 1. Mode : | Mode mengindikasikan modality (mayor atau minor) dari suatu track, tipe dari tangga nada yang menurunkan melodi |
| 1. Name : | Nama |
| 1. Popularity : | Popularitas dari track dengan nilai antara 0 dan 100, dengan 100 adalah paling popular. Dihitung oleh algoritma dan didasarkan pada total jumlah suatu track diputar dan pemutaran terkini. |
| 1. Release Data : | Tahun perilisan |
| 1. Speechiness : | Mendeteksi keberadaan kata-kata dalam suatu track. |
| 1. Tempo : | Tempo dari suatu track dalam ukuran beats per minute (BPM). |

Data diambil dari Kaggle dan algoritma yang digunakan untuk prediksi adalah Random Forest.

**Key Flows**

Project pipeline yang digunakan adalah end to end machine learning hingga deployment model. Struktur data dari Project disesuaikan dengan best practice dalam pembuatan machine learning di mana terdapat beberapa script phyton utama yang dijalankan yaitu :

1. data\_pipeline.py : berisikan modul terkait data collection, data validation
2. preprocessing.py : berisikan modul terkait handling missing values
3. Modeling : modul terkait baseline model, evaluation
4. Pytest : modul pengetesan coding python
5. Api.py dan Streamlit.Py : modul terkait API sebagai penghubung dengan user dan streamlit sebagai user interface.

Untuk service API dan streamlit ditunjang menggunakan docker service sebagai virtual environment

**Launch Readiness**

Project dari timeline sesuai dengan deadline dari pengerjaan tugas proyek ML Process tanggal 9 November 2024

**Artifact**

Komponen yang digunakan dalam proyek ini yaitu Visual Studio Code, Jupyter Notebook, Google Chrome User, Git Hub, Git Bash, Docker. Adapun modul python yang digunakan adalah pandas, numpy, sklearn, matplotlib, seaborn, dan seterusnya

**References**

Spotify Data Analysis and Song Popularity Prediction : Sivasai Bhavanasi, Sahil Malla, V Manichetan, CVNJ Dhanush, Dr B Prakash

<https://www.kaggle.com/datasets/leonardopena/top-spotify-songs-from-20102019-by-year>

**Source Code end to end workflow**

1. Exploratory Data Analysis

Tahapan ini berada pada file Jupyter Notebook untuk melakukan analisis yang nantinya akan diubah ke dalam bentuk \*py python file

* 1. Cek ketersediaan predictor

Fitur yang digunakan sebagai predictor adalah accousticness, danceability, duration, energy, loudness, speechiness, valence

* 1. Cek data types

Saat pengecekan awal, data types untuk semua fitur merupakan float

* 1. Cek data hilang, NaN, dan Outliers

Tidak ditemukan data yang hilang.

* 1. Cek Range dari data

Pengecekan range data dilakukan untuk mendapatkan batas bawah dan batas atas dari data.

* 1. Pengecekan fitur lainnya

1. Data Preprocessing
   1. Analisa Univariate

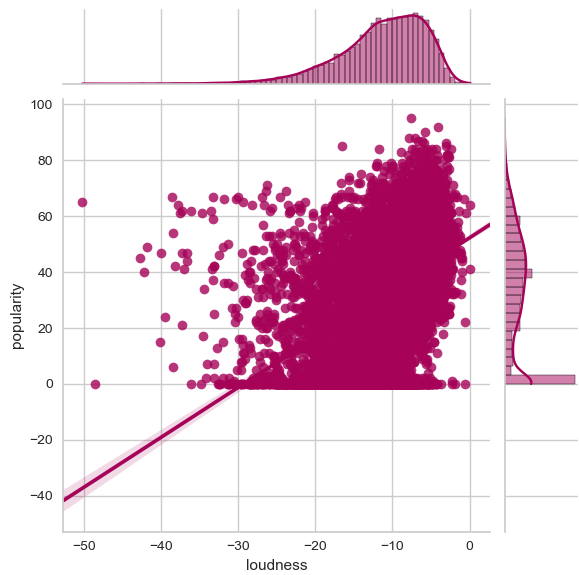
Dilakukan analisa Univariate untuk menghandle outlier. Untuk itu data distandardisasi ke mean 0 dan standard deviasi 1

* 1. Analisa bivariate
* Didapatkan hasil bahwa loudness memiliki relasi yang baik dengan popularity,
* classical piano" genres memiliki "acousticness" dan tidak memiliki "speechiness" dan "energy"
* movie tunes" dan "show Tunnes" memiliki karakteristik Audio "valence" yang rendah dan "acousticness" yang tinggi.

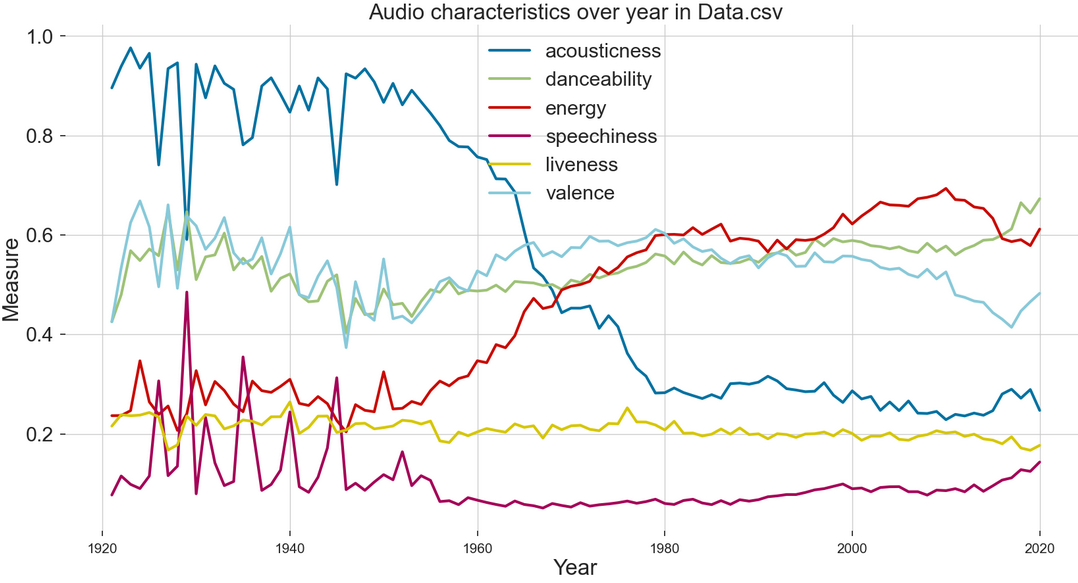
A comparison of a graph

Description automatically generated with medium confidence

loudness memiliki relasi yang baik dengan popularitas, berdasarkan genre, karakteristik audio akan berubah sesuai dengan key yang digunakan. sebagian besar genre "acousticness" memiliki efek lebih "movie tunes" and "show Tunnes" terlihat memiliki karakteristik audio "valence" rendah dan acousticness tinggi genre "classical piano" memilki "acousticness" dan tidak ada "speechiness" and "energy"



Perubahan karakteristik audio sejalan dengan waktu untuk berbagai variabel



1. Feature Selection

Dari analisa korelasi, yang memiliki korelasi positif dengan dependent variable adalah key, durasi, tempo, loudness, energy dan danceability. Sedangkan yang memiliki korelasi negative adalah mode, valance, speechiness, liveness, serta accousticness.

A graph of a bar graph

Description automatically generated

1. Modeling

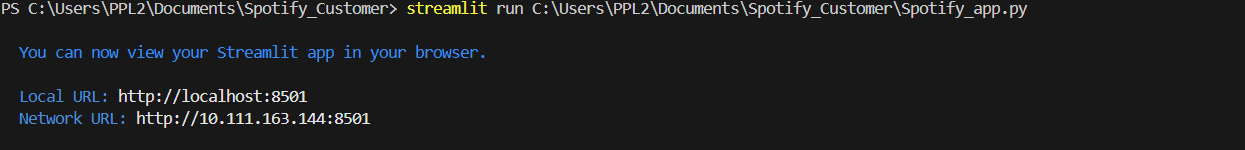
Model yang digunakan adalah Random Forest. Hasil dari model dapat dievaluasi dengan menggunakan metrics accuracy adalah sebesar 93.68.

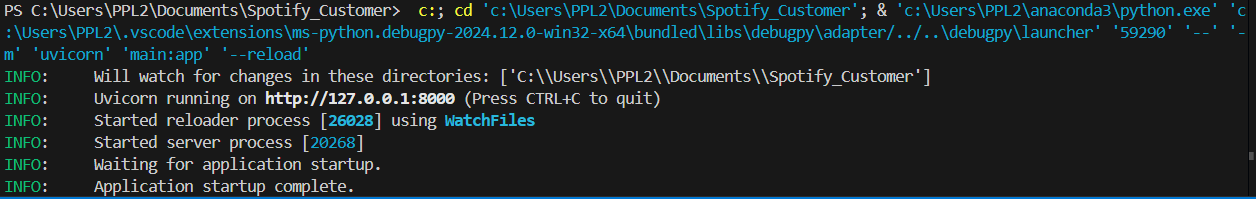
1. Serving

Setelah pipeline dari data pipeline, preprocessing dan modeling didapatkan dan dilakukan pengetesan, maka dilakukan serving deployment model. Ada 2 servis penting yang digunakan dalam deployment model yaitu Fast API dan Streamlit. Fast API digunakan sebagai jalur komunikasi antara Front End dan Back End sehingga data yang didapat dari user bisa digunakan untuk predict data menggunakan Fast API.

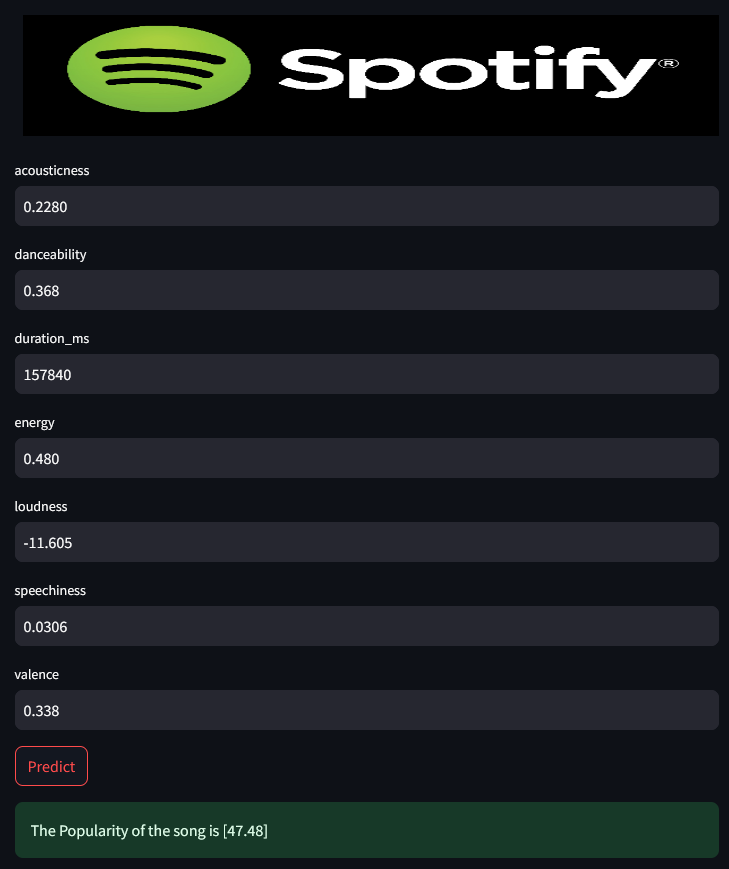
File dari main.py berisikan modul seperti Fast API, base model yang merupakan object class dari pydantic yang inherit dari base model, uvicorn untuk web server yang digunakan oleh python dan file python berupa pipeline yang telah dibuat. Setelah pembuatan object FAST API, dibuat struktur input data dan memuat model menggunakan class.

Streamlit digunakan sebagai User Interface agar user dapat memasukkan data. File streamlit.py berisikan modul streamlit, request yang digunakan untuk melakukan request pada web yang mengizinkan untuk mengirimkan HTTP request, dan PIL image untuk memasukkan gambar. Selanjutnya dibuat inputan dan select box menggunakan streamlite dan dibuat submit button berupa “Predict”. Hasil input tersebut dimasukkan ke dalam JSON file dan diberikan output akhir berupa prediksi popularitas. Cara untuk melakukan akses adalah mengeksekusi pada terminal untuk file api dan menjalankan streamlit pada terminal secara terpisah.



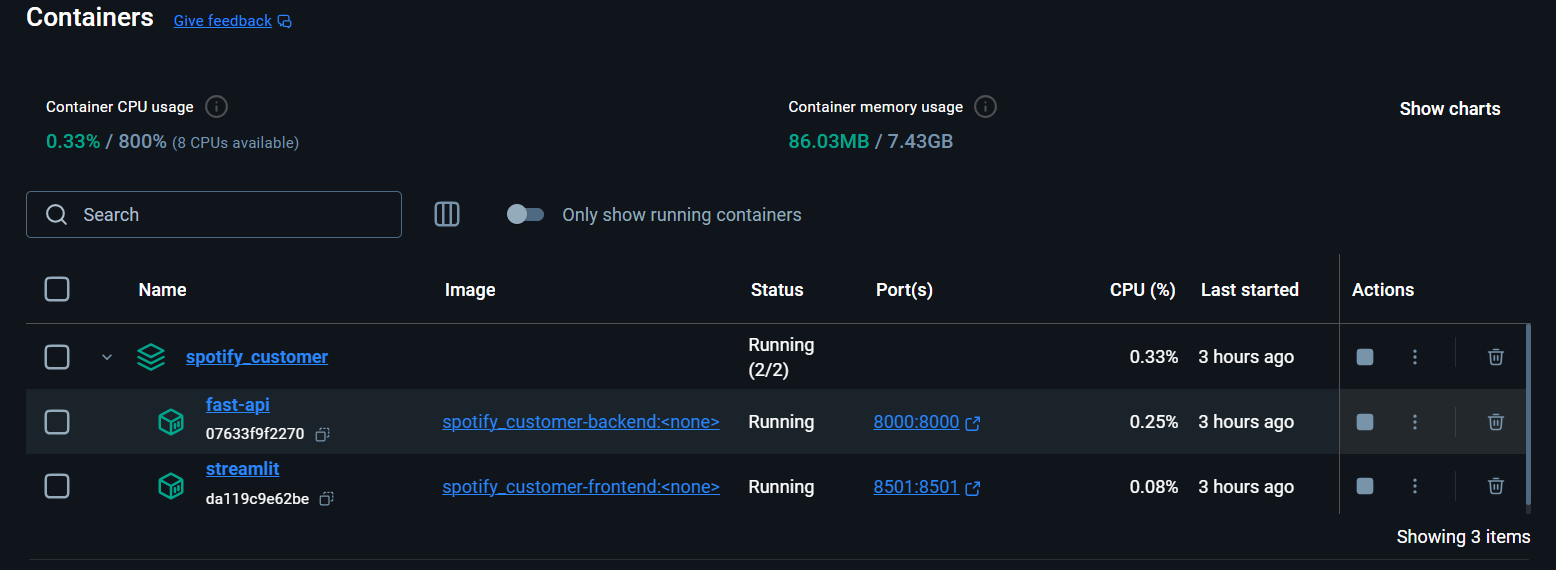


Hasil interface dari servis yang digunakan untuk prediksi popularitas lagu pada Spotify:

****

Sebagai langkah pertama, docker di build terlebih dahulu sehingga menghasilkan docker image yang selanjutnya dilakukkan run pada wsl/ubuntu. Dua buah docker file tersebut menggunakan python:3.9.15-slim-buster dan terdapat dua docker image untuk masing-masing servis. Docker akan melakukan install terhadap dependensi yang telah ditulis pada requirement.txt.

Agar kedua docker dapat berjalan secara bersama, maka digunakan docker-composs.yaml. File ini berisikan services yang digunakan (streamlit dan api), yang terdapat konfigurasi nama terhadap build, image, container\_name, ports, dan volume (directory).

****

Sampai tahapan ini, service sudah dapat berjalan pada localhost. Selanjutnya dilakukan deployment model secara online agar dapat diakses oleh user. Deployment dilakukan menggunakan server dari AWS menggunakan Instance pada EC2. Untuk menghubungkannya, seluruh file pipeline machine learning tersebut dimasukkan ke Git Hub. Untuk memasukkan ke Git Hub, dilakukan dengan membuat repository GitHub terlebih dahulu secara online.

Setelah repository dibuat, maka dilakukan “git init” untuk menginisiasi github pada folder working directory. Selanjutnnya dilakukan “git add .” untuk track perubahan dokumen yang nantinya akan dilakukan “git commit”. Setelah dilakukan commit maka seluruh pipeline project dapat di push pada github. Selanjutnya untuk membuat deployment yang otomatis dapat digunakan CICD menggunakan github action dengan melakukan setting pada git-hub action