LAPORAN KECERDASAN KOMPUTASIONAL – IF184503

FINAL PROJECT KECERDASAN KOMPUTASIONAL

Kelompok 6 :

Alie Husaini R.
05111840000097

Ammar Alifian
05111184000007

Nodas Uziel Putra Serpara
05111840007007

ABSTRAK

Untuk menganalisis kecocokan lagu dan klasifikasinya untuk era tertentu, kami melakukan analisis data dengan dataset lagu-lagu di Spotify dari tahun 1920 - 2020. Pertama kami menganalisis artis terpopuler untuk tiap dekade, lalu mengamati fitur-fitur apa yang menjadi faktor terpenting dalam popularitas suatu lagu. Dengan menggunakan SelectKBest, kami mendapati bahwa acousticness, energy, dan loudness adalah faktor terpenting untuk popularitas lagu. Selain itu kami melakukan clustering untuk mengelompokkan lagu-lagu di Spotify. Hasilnya, lagu terbagi menjadi 4 cluster besar. Kami juga mengklasifikasi lagu berdasarkan era atau dekade, dan didapati bahwa klasifikasi terbaik adalah dengan metode Decision Tree dan MLP. Namun, dengan tingkat akurasi yang relatif rendah dapat disimpulkan bahwa antar dekade variasi lagu yang ada sama luasnya.

Kata Kunci: Clustering, Eksplorasi Data, Klasifikasi, SelectKBest, Spotify

DAFTAR ISI

BAB 1.
PENDAHULUAN

1.1. Latar Belakang

Laporan ini dibuat untuk memenuhi tugas akhir mata kuliah Kecerdasan Komputasional. Dataset yang digunakan berasal dari website Kaggle tentang data lagu di Spotify dari tahun 1921-2020 (https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks?select=data.csv). data tersebut memuat tentang lagu-lagu yang ada semenjak tahun 1921-2020, dengan berbagai macam fitur audio yang dapat mempengaruhi popularitas suatu lagu. Output pertama yang ingin dicapai adalah mengetahui penyebab populernya suatu lagu pada suatu era, apakah lagu-lagu dulu memiliki popularitas yang lebih kecil dari lagu-lagu sekarang? Apakah fitur audio yang paling berkontribusi pada popularitas lagu? Untuk output kedua yang ingin dicapai yaitu mengetahui berapa banyak pengelompokan lagu yang dapat dilakukan pada data Spotify. Asal usul output kedua ini berdasarkan salah satu data yang disediakan di website Kaggle pada dataset Spotify dari tahun 1921-2020, yaitu dataset lagu yang memuat genre lagu tersebut, namun dikarenakan genre lagu yang ada berjumlah lebih dari 5000 sehingga kurang dapat diandalkan sebagai pengelompokan lagu-lagu yang ada. Kami pun memikirkan cara apakah yang dapat dipakai dalam mengelompokkan lagu-lagu yang ada pada data Spotify tersebut.

1.2. Perumusan Masalah

Masalah yang dirumuskan adalah sebagai berikut :

1. Artis terpopuler tiap dekade?
2. Fitur apa yang sangat mempengaruhi popularitas suatu lagu?
3. Data lagu Spotify dari tahun 1921-2020 dapat dikategorikan menjadi berapa kelompok?
4. Model klasifikasi apa yang paling bagus untuk mengklasifikasi lagu berdasarkan era masing-masing?

BAB 2.

DESAIN DAN IMPLEMENTASI

2.1. Persiapan Data

Untuk Final Project ini, kami menggunakan dataset dari Kaggle untuk menunjukkan data mendetail dari lagu-lagu yang ada di Spotify[1]. Data-data yang ada dalam dataset tersebut adalah :

- Data numerical, terdiri dari :
    - acousticness (Representasi dari seberapa akustik lagu tersebut, bernilai dari 0 hingga 1)
    - danceability (Kelayakan lagu tersebut untuk digunakan sebagai pengiring tarian, bernilai dari 0 hingga 1)
    - energy (Energi yang terpancar dari lagu tersebut, bernilai dari 0 hingga 1)
    - duration_ms (Durasi lagu dalam milisecond)
    - instrumentalness (Instrumentalitas dalam lagu tersebut, bernilai dari 0 hingga 1)
    - valence (Positivitas yang terpancar dari lagu tersebut, bernilai dari 0 hingga 1)
    - popularity (Seberapa populer lagu tersebut, bernilai dari 0 hingga 100)
    - tempo (Tempo lagu dalam BPM)
    - liveness (Keberadaan penonton, bernilai dari 0 hingga 1)
    - loudness (Kerasnya lagu rata-rata dalam dB)
    - speechiness (Frekuensi lirik dalam lagu, bernilai dari 0 hingga 1)
    - year (Tahun rilis lagu, berkisar dari 1921 hingga 2020)
- Dummy, terdiri dari :
    - mode (Tipe kunci lagu, 0 = Minor, 1 = Major)
    - explicit (Sifat suatu lagu, apakah eksplisit (mengandung umpatan, unsur seksual, narkoba, dsb) atau tidak. 0 = Tidak eksplisit, 1 = Eksplisit)
- Categorical, terdiri dari :

- key (Kunci lagu yang dikodekan dengan integer, berkisar dari 0 - 11, kunci C = 0, C# = 1, dan seterusnya)
- artists (Daftar artis dari lagu tersebut)
- release_date (Tanggal rilis lagu, dalam format yyyy-mm-dd)
- name (Judul lagu)

Selain itu, untuk permasalahan nomor 4 kami menggunakan target yaitu era atau dekade. Target tersebut diambil dari fitur tahun, kemudian digolongkan menjadi 1920an, 1930an, dan seterusnya.

2.2. Skenario Uji Coba

Untuk penyelesaian masalah nomor 1, digunakan pencarian dengan cara mencari lagu terpopuler dari masing-masing dekade, lalu kami mengambil nama artist dari lagu tersebut.

Untuk penyelesaian masalah nomor 2, digunakan *library* SKLearn dan menggunakan fungsi SelectKBest pada SKLearn untuk menentukan keberpengaruhan suatu aspek pada popularitas lagu.

Untuk penyelesaian masalah nomor 3, digunakan Principal Component Analysis (PCA) dan K-means untuk membagi tiap Playlist (Cluster) berdasarkan keunggulan fitur audio masing-masing Playlist.

Untuk penyelesaian masalah nomor 4, digunakan model SVM, KNN, MLP, dan Decision Tree dalam mengklasifikasi. dan dibandingkan hasilnya untuk mencari klasifikasi yang terbaik untuk dataset Spotify.

BAB 3.

HASIL UJI COBA DAN DISKUSI

3.1. Hasil Uji Coba Skenario

3.1.1. Artis terpopuler tiap dekade?

Dalam menjawab masalah ini, dilakukan pencarian artis dengan nilai popularitas tertinggi pada tiap 1 dekade dari 1921-2020. Karena fitur popularity melekat pada lagu dan bukan artist, kami memutuskan bahwa untuk setiap artist, popularity dari artist tersebut sejumlah max(popularity) dari semua lagu dalam suatu dekade oleh artist tersebut. Dan hasil yang didapat adalah :

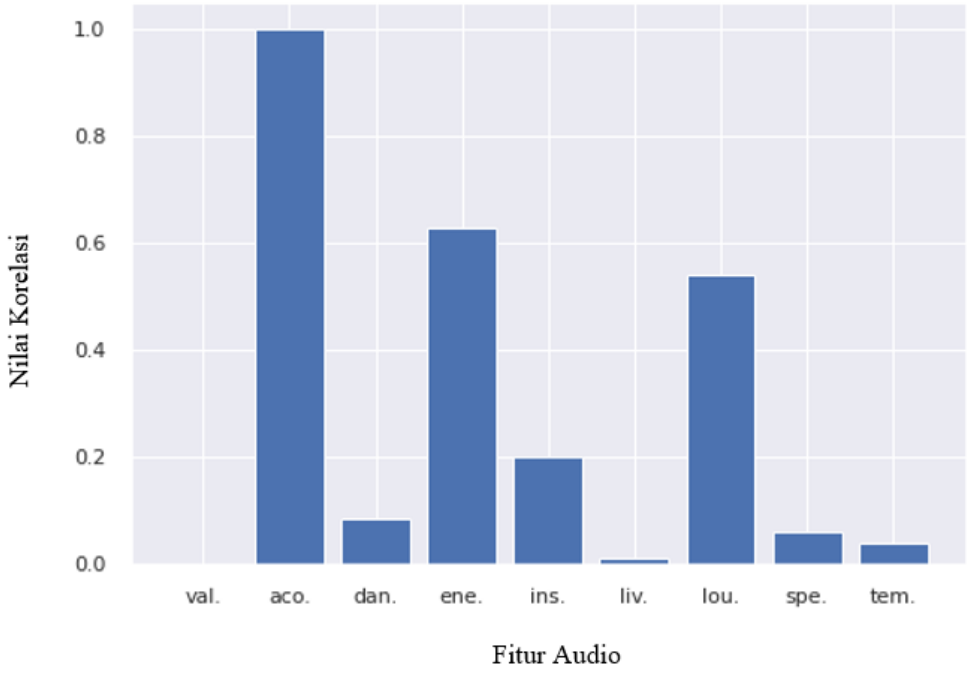
Nama Artis	Popularitas	Tahun
Louis Armstrong	52	1921-1930
Billie Holiday, Eddie Heywood	64	1930-1939
Bing Crosby, Ken Darby Singers, John Scott Trotter & His Orchestra	76	1940-1949
Dean Martin	81	1950-1959
Brenda Lee	85	1960-1969
Fleetwood Mac	89	1970-1979
AC/DC	84	1980-1989
The Police	84	1980-1989
a-ha	84	1980-1989
Mariah Carey	88	1990-1999
Coldplay	84	2000-2009
Linkin Park	84	2000-2009
Eminem, Nate Dogg	84	2000-2009
Harry Styles	94	2010-2019

Tabel 1. Artis terpopuler tiap dekade

Dari Tabel 1 dapat disimpulkan bahwa terdapat beberapa nama artis yang memiliki popularitas tinggi berdasarkan lagu yang dinyanyikan bersama atau lagu yang diciptakan berdasarkan kolaborasi di antara 2 atau lebih artis. seperti Billie Holiday dan Eddie Heywood dengan judul lagu *Wherever you are (Live)*. Selain itu, lagu-lagu dalam beberapa dekade terakhir cenderung lebih populer dibandingkan lagu-lagu dari awal abad ke-20.

3.1.2. *Fitur apa yang sangat mempengaruhi popularitas suatu lagu?*

Dalam menjawab masalah ini, dilakukan menggunakan Sklearn dan SelectKBest, membuat Data Train dan Data Test, lalu membuat function yang mengecek seberapa pengaruh tiap fitur lagu yang ada. Jika divisualisasikan maka akan menjadi seperti Gambar 2.

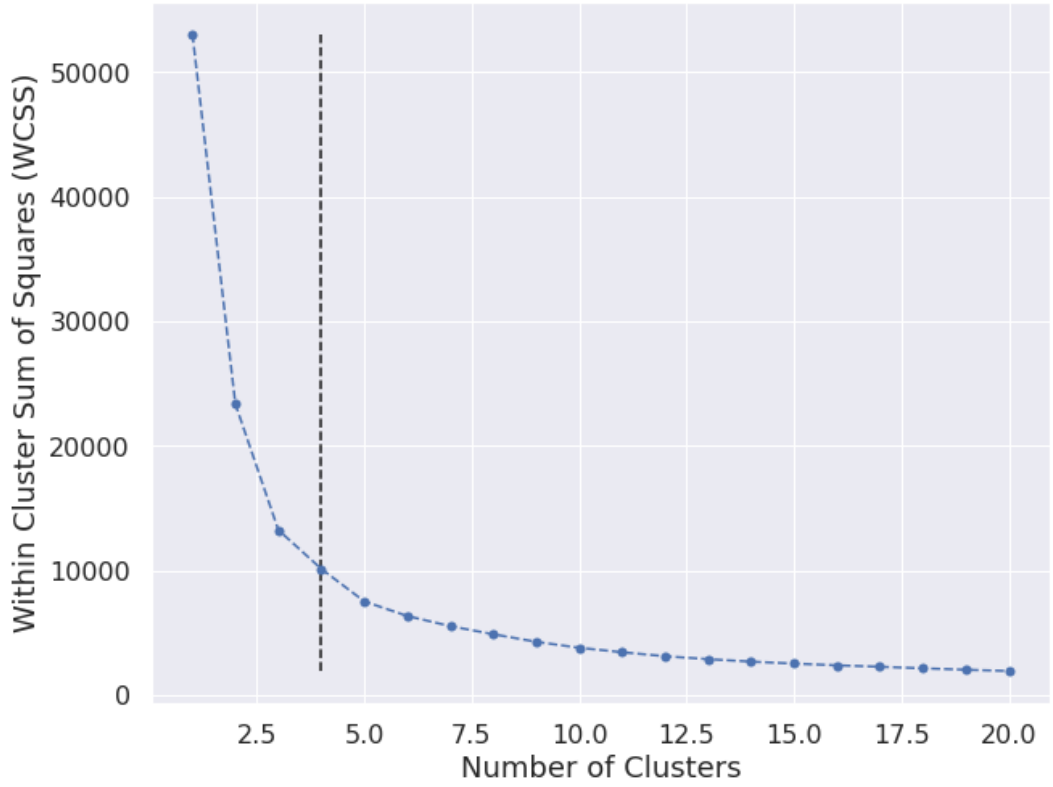


Gambar 2. Histogram Fitur Audio yang Mempengaruhi Nilai Popularitas Suatu Lagu

Dapat disimpulkan dari Gambar 2. bahwa fitur audio yang paling mempengaruhi popularitas suatu lagu adalah *acousticness*. dan 2 diantaranya yang juga mempengaruhi adalah fitur audio *energy* dan *loudness*.

3.1.3. *Data lagu Spotify dari tahun 1921-2020 dapat dikategorikan menjadi berapa kelompok?*

Dalam menjawab masalah ini, digunakan MinMaxScaler untuk normalisasi dan Principal Component Analysis (PCA) untuk proyeksi pada bidang 2 dimensi, lalu dicari optimal cluster yang terbaik untuk mengelompokan lagu-lagu tersebut. divisualisasikan seperti Gambar 3:

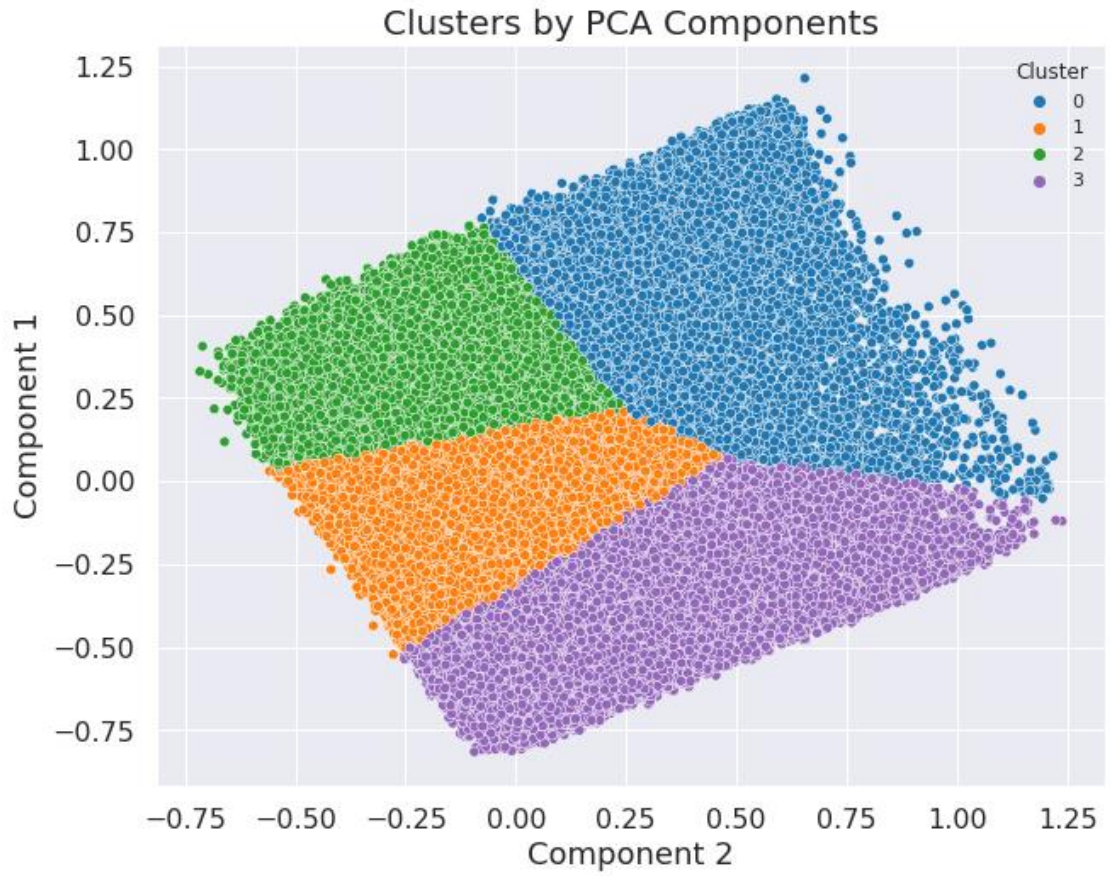


Gambar 3. Visualisasi Cluster yang Optimal

Sesuai Gambar 3. bahwa jumlah Cluster yang paling optimal dalam mengelompokan lagu-lagu dalam Spotify adalah 4 Cluster. Setelah itu dipakailah K-means untuk mengclustering lagu-lagu yang ada berdasarkan fitur audionya dan membentuk 2 Component yang membantu pengelompokan 1 lagu dalam suatu cluster. Jika dilihat pada Tabel 2. terjadi pengelompokan 4 kategori lagu berdasarkan fitur audio dan Component pada lagu tersebut lalu Gambar 4. adalah bentuk visualisasi scatterplot pada masing-masing lagu yang sudah dikelompokan berdasarkan fitur audio dan Componentnya.

	valence	acousticness	danceability	energy	instrumentalness	liveness	loudness	speechiness	tempo	Component 1	Component 2	Cluster
0	0.0594	0.985944	0.282389	0.211	0.878000	0.665	0.624916	0.037732	0.332450	0.914791	0.521781	0
1	0.9630	0.734940	0.828947	0.341	0.000000	0.160	0.744797	0.427835	0.250243	0.038439	-0.532255	1
2	0.0394	0.964859	0.331984	0.166	0.913000	0.101	0.707071	0.034948	0.453125	0.925289	0.541565	0
3	0.1650	0.970884	0.278340	0.309	0.000028	0.381	0.793736	0.036495	0.411113	0.484185	-0.143180	2
4	0.2530	0.960843	0.423077	0.193	0.000002	0.229	0.781521	0.039175	0.417503	0.493923	-0.234173	2

Tabel 2. Pembagian Lagu berdasarkan Fitur Audio



Gambar 4. Scatter Plot Kelompok Lagu(Cluster)

Cluster	valence	acousticness	danceability	energy	instrumentalness	liveness	loudness	speechiness	tempo	Component 1	Component 2
0	0.352199	0.891423	0.401128	0.230467	0.813454	0.182372	0.657700	0.057572	0.434916	0.725275	0.371140
1	0.633379	0.472830	0.611914	0.504980	0.021001	0.206679	0.773925	0.133618	0.486537	-0.126280	-0.151621
2	0.470059	0.878609	0.509080	0.258469	0.053613	0.207210	0.720335	0.113340	0.455991	0.362983	-0.259969
3	0.576573	0.080463	0.584271	0.741501	0.084894	0.213845	0.823082	0.089159	0.512222	-0.490817	0.146087

Tabel 3. Fitur Audio pada masing-masing kelompok (Cluster)

Dapat disimpulkan bahwa, dapat dilakukan pembagian kelompok lagu sebanyak 4 kelompok (Cluster). Dan setiap kelompok memiliki keunggulan fitur audio tersendiri, seperti Tabel 3. Terlihat bahwa Cluster 0 memiliki fitur audio acousticness dan *instrumentalness* paling tinggi diantara cluster yang lain. Cluster 1 memiliki nilai yang relatif sedang untuk setiap fitur audio. Cluster 2 memiliki nilai acousticness yang hampir setinggi cluster 0, namun nilai instrumentalness justru berkebalikan dengan cluster 0. Cluster 3 memiliki nilai acousticness yang sangat rendah, namun juga nilai energy yang cukup tinggi dibandingkan cluster-cluster lain.

3.1.4. Model klasifikasi apa yang paling bagus untuk mengklasifikasi lagu berdasarkan era masing-masing?

Dalam menjawab masalah ini, digunakan klasifikasi model SVM, MLP, Decision Tree, dan KNN. tetapi dalam percobaan SVM terdapat kendala waktu yang sangat lama dalam mencari akurasi, walaupun sudah menggunakan 5% dari dataset Spotify, proses pencarian yang dilakukan berlangsung sampai 1 setengah jam lebih, dan tidak membuahkan hasil apapun, sehingga SVM bukan model yang cocok untuk mengklasifikasi data ini. Dan dari percobaan model Decision tree, MLP, dan KNN, klasifikasi terbaik dalam dataset Spotify adalah Model Decision Tree dan MLP. Berikut gambar uji coba yang dilakukan :

Classifier	Test Size	Criterion	Splitter	Max Depth	Accuracy
Decision Tree	30%	Gini	Best	3	0.303
				5	0.306
				7	0.327
				9	0.348
			Random	7	0.3
				9	0.317
		Entropy	Best	7	0.328
				9	0.347
	50%	Gini	Best		0.344
		Entropy			0.339
	70%	Gini			0.338
		Entropy			0.336

Tabel 4. Hasil percobaan dengan menggunakan Decision Tree

Classifier	Test Size	Weights	N Neighbors	Accuracy
K-Nearest Neighbor	30%	Uniform	1	0.232
			11	0.248
			21	0.249
			31	0.248
		Distance	21	0.265
			31	0.265
	50%		21	0.283
			31	0.283
	70%		21	0.283
			31	0.283

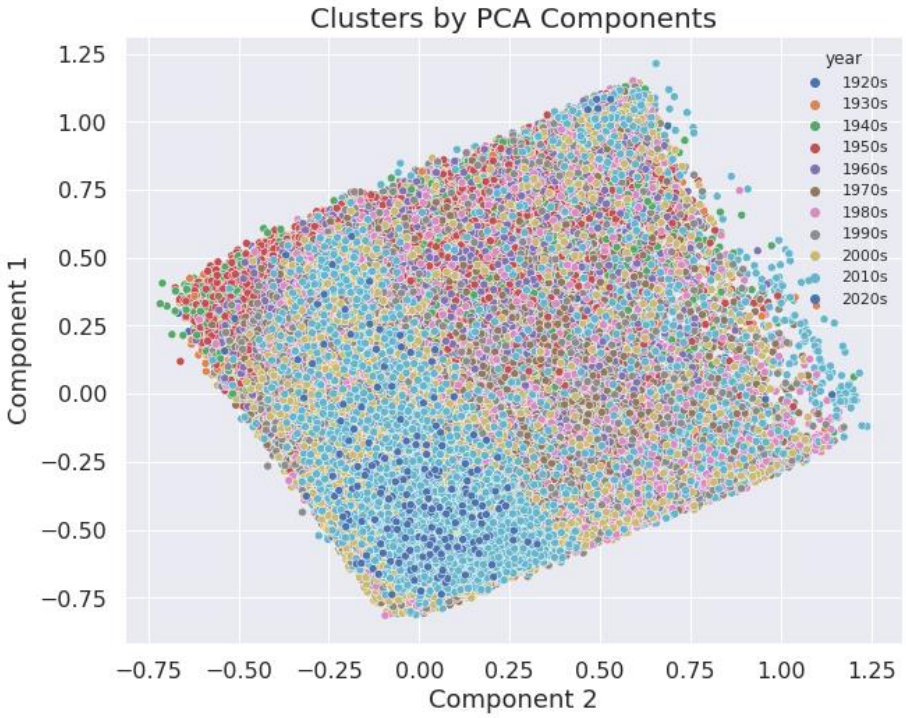
Tabel 5. Hasil percobaan dengan menggunakan KNN

Classifier	Test Size	Solver	Activation	Alpha	Accuracy
Multilayer Perceptron	30%	Adam	ReLU	0.0001	0.345
				0.001	0.352
				0.01	0.333
				0.1	0.328
				1	0.302
			Identity	0.0001	0.313
				0.001	0.31
			Logistic	0.0001	0.348
				0.001	0.341
			Tanh	0.0001	0.34
				0.001	0.338
		SGD	ReLU	0.0001	0.256
			Logistic		0.268
	50%	Adam	ReLU		0.338
			Logistic		0.329
	70%		ReLU		0.335
			Logistic		0.328

Tabel 6. Hasil percobaan dengan menggunakan MLP

Dengan melihat data dari Tabel 4, Tabel 5, dan Tabel 6, bisa dilihat bahwa percobaan dengan Decision Tree dan MLP memiliki akurasi paling baik yaitu 0.348 atau 34.8%. Sementara percobaan dengan kNN memiliki akurasi 0.283 atau 28.3%.

Dari hasil uji coba skenario 4, rata-rata akurasinya 35% ke bawah, bisa jadi dikarenakan rentang era yang kami gunakan adalah per dekade sehingga ada cukup banyak keragaman musik untuk setiap era.



Gambar 6. Scatter Plot lagu sesuai tahun rilisnya

Dikarenakan akurasi yang sangat rendah pada tiap metode klasifikasi yang dilakukan, kami mengambil kesimpulan bahwa antar dekade, variasi lagu yang ada sama luasnya sehingga sulit untuk mengklasifikasikan lagu berdasarkan dekade dengan tingkat akurasi yang tinggi. Untuk mendukung kesimpulan ini, dilakukan uji coba clustering berdasarkan dekade menggunakan component yang didapat pada masalah nomor 3 (*Data lagu Spotify dari tahun 1921-2020 dapat dikategorikan menjadi berapa kelompok?).* Pada Gambar 6, terlihat bahwa antar dekade, setiap lagu relatif menyebar dan terjadi overlapping. Setelah dilakukan klasifikasi menggunakan sumbu x dan y komponen PCA pun didapat bahwa akurasi semakin berkurang. Selain itu, dari beberapa kali running yang sukses dengan menggunakan SVM, akurasi juga relatif sama, sehingga jelas bahwa masalah bukan pada model klasifikasi. Kami juga menemukan beberapa sumber yang mendefinisikan bahwasannya musik sejak tahun 1930an dan selanjutnya dikategorikan sebagai satu era yaitu postmodern, sehingga tidak mengejutkan jika lagu-lagu abad ke-20 memang punya variasi yang begitu luas.

3.2.Kesimpulan

Dari uji coba diatas, kami menyimpulkan bahwa :

1. Artis terpopuler untuk tiap dekade adalah seperti yang tercantum dalam Tabel 1.
2. Fitur audio yang paling menentukan popularitas suatu lagu adalah *acousticness*, *energy*, dan *loudness*.
3. Dalam pembagian yang dilakukan dihasilkan pengelompokan yang paling optimal dengan jumlah kelompok sebanyak 4. Dan tiap kelompok memiliki keunggulan fitur audio masing-masing.
4. Metode klasifikasi terbaik per era adalah dengan metode Decision Tree dan MLP.
5. Klasifikasi lagu berdasarkan era atau dekade tidak dapat memperoleh akurasi yang tinggi dikarenakan luasnya variasi lagu untuk setiap dekade.

DAFTAR PUSTAKA

[1] Ay, Yamaç Eren (2020). Spotify Dataset 1921-2020, 160k+ Tracks. https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks?select=data.csv

[2] Karolyi, Otto. 1994. Modern British Music: The Second British Musical Renaissance—From Elgar to P. Maxwell Davies. Rutherford, Madison, Teaneck: Farleigh Dickinson University Press; London and Toronto: Associated University Presses. ISBN 0-8386-3532-6

[3] Meyer, Leonard B. 1994. Music, the Arts, and Ideas: Patterns and Predictions in Twentieth-Century Culture, second edition. Chicago and London: University of Chicago Press. ISBN 0-226-52143-5.

`