Analisa Audio Features dengan Membandingkan Metode Multiple Regression dan Polynomial Regression untuk Memprediksi Popularitas Lagu

Billy Faith Susanto[1], Silvia Rostianingsih[2], Leo Willyanto Santoso[3]
Program Studi Informatika, Fakultas Teknologi Industri, Universitas Kristen Petra
Jl. Siwalankerto 121-131, Surabaya 60236
Telp (+6231) – 2983455, Fax. (+6231) - 8417658
billysusanto.bs@gmail.com[1] , silvia@petra.ac.id[2], leow@petra.ac.id[3]

ABSTRAK

Lagu merupakan karya seni yang mengeskpresikan ide dan emosi dalam bentuk ritme, melodi, harmoni. Lagu menjadi sumber keuntungan yang besar bagi musisi atau artis melalui sisi komersialnya. Berdasarkan data dari IFPI, pendapatan dari industri musik pada tahun 2019 sebesar 20.2 milliar USD, di mana 56.1% didapat dari streaming revenue. *Spotify* merupakan salah satu *streaming service* paling populer di dunia.

Penelitian ini mencoba untuk melakukan prediksi popularitas dari sebuah lagu berdasarkan data *audio features*-nya yang diambil dari *Spotify* API. Dalam melakukan prediksi akan digunakan 2 metode regresi, yaitu *Linear Regression* dan *Polynomial Regression*. Model yang dibuat menggunakan metode tersebut akan diukur akurasinya dengan metrik R2, Adjusted R2, MAE, dan MSE.

Berdasarkan hasil analisis pengimplementasian program, metode *Linear Regression* mempunyai hasil R2 rata-rata sebesar 0.23614, *Adjusted* R2 rata-rata sebesar 0.23536, dan memiliki error dengan metode MAE rata–rata sebesar 17.38129, MSE rata–rata sebesar 442.31700. Metode *Polynomial Regression* mempunyai hasil R2 rata-rata sebesar 0.31496, Adjusted R2 rata-rata sebesar 0.25880, dan memiliki error dengan MAE rata–rata sebesar 16.47367, MSE rata–rata sebesar 409.76242.

Kata Kunci: *regresi linear, regresi polinomial, prediksi popularitas, fitur audio Spotify.*

ABSTRACT

Songs are artistic works that expresses ideas and emotion in the forms of rhythms, melodies, and harmonies. Songs are the source of huge profit for musicians or artists from commercial view-point. Based on the data from IFPI, the earnings from the music industry in 2019 reached US$20.2 billion, in which 56.1% of them came from streaming revenue. Spotify is one of the largest and most well-known streaming services in the world today.

This research aims to make predictions of popularity from each song according to the audio feature data taken from Spotify's API. The process of prediction will use 2 regression methods, which are Linear Regression and Polynomial Regression. The model will be made using those 2 methods and will be tested with the R2, Adjusted R2, MAE, and MSE metric systems.

From the analysis of the implementation to the program, the Linear Regression method had garnered the average results as follows: 0.23614 for R2, 0.23536 for Adjusted R2, and had average errors 17.38129 for MAE method, 442.31700 for MSE method. Using the Polynomial Regression method, the average results were: 0.31496 for R2, 0.25880 for Adjusted R2, and had average errors 16.47367 for MAE method, 409.76242 for MSE method.

Keywords: *linear regression, polynomial regression, popularity predictor, Spotify audio features.*

1. PENDAHULUAN

1.1 Latar Belakang

Lagu merupakan komposisi musikal yang dimaksudkan untuk dinyanyikan dengan kata-kata. Musik merupakan seni suara yang mengeskpresikan ide dan emosi dalam bentuk ritme, melodi, harmoni, dan warna. *Music Track* atau biasanya disebut sebagai

Produksi musik sekarang dapat dilakukan sendiri tanpa perlu ke studio profesional dikarenakan aksesibilitas ke alat rekaman sudah menjadi lebih mudah dan terjangkau. Artis atau musisi baru pun dapat menciptakan lagu dan kemudian memasarkannya untuk tujuan komersial. *Spotify* merupakan layanan *music streaming* dengan jumlah pengguna paling besar di dunia [2]. Artis dapat meng-upload *track* musik karyanya ke *Spotify* dan kemudian bersaing dengan lagu lainnya untuk mendapatkan semakin banyak jumlah pendengar.

Di sisi lain dalam industri ini, *record labels* mengeluarkan biaya yang sangat besar untuk memproduksi lagu dan mempromosikan artis baru. Menurut *International Federation of the Phonographic Industry* (IFPI), setiap tahunnya sebanyak 5.8 miliar USD diinvestasikan dalam A&R (*Artists and Repertoire*) dan *Marketing* oleh *record labels*. Pendapatan dari industri musik sebesar 20.2 milliar USD, di mana 56.1% didapat dari *streaming revenue* [9].

Musik memang merupakan karya seni yang membutuhkan kreatifitas. Tetapi untuk tujuan komersialnya, kreatifitas harus diikuti dengan pemahaman akan trend yang disukai oleh pasar, yaitu masyarakat sebagai pendengarnya. Sesulit dan secanggih apapun proses produksinya, tetapi jika tidak sesuai dengan keinginan pasar, maka secara komersial lagu tersebut menjadi gagal. Biaya yang diinvestasikan pun menjadi sia-sia. Penelitian ini akan memprediksi popularitas lagu berdasarkan faktor teknis apa yang ada di dalam lagu tersebut.

Setiap *track* yang ada di *platform Spotify* mempunyai data *Audio Features*-nya yang sudah dianalisa oleh algoritma *Spotify* sendiri [16]. *Audio features* tersebut mendeskripsikan nilai-nilai teknis dari lagu, yaitu *duration* (durasi dalam satuan milliseconds), *key* (estimasi nada dasar), *mode* (mayor atau minor), *danceability* (seberapa cocok lagu tersebut untuk *dance*), *instrumentalness* (memprediksi apakah lagu tersebut tidak mengandung vokal), *liveness* (mendeteksi apakah ada kehadiran penonton dalam rekaman lagu yang menandakan lagu mungkin direkam secara

live), *speechiness* (mendeteksi keberadaan kata-kata yang diucapkan), *valence* (menjelaskan tentang seberapa positif lagu tersebut tersampaikan, yaitu *happy*, *cheerful*, *euphoric*), tempo (estimasi rata-rata tempo dalam BPM / *beats per minutes*), *loudness* (seberapa kuat track dalam satuan decibels), *acousticness* (*confidence measure* apakah track tersebut akustik), dan *energy* (*perceptual measure*, track terasa cepat, keras, dan ribut). Data yang terkandung ada dalam bentuk *numerical value* dan ada dalam bentuk *categorical value*. Nilai *popularity* yang terdapat dalam dataset *Spotify* dihitung oleh algoritma *Spotify* berdasarkan *total number of plays* dan seberapa baru *plays* tersebut.

Hit Song Science (HSS) merupakan sebuah topik saintifik yang membahas tentang prediksi lagu yang akan menjadi populer berdasarkan *audio features*-nya. Penelitian yang dilakukan oleh Herremans et al. membuktikan bahwa *audio features* mempunyai performa yang bagus dalam memprediksi apakah suatu lagu merupakan *'top 10' dance hit* dibandingkan dengan lagu lain yang berada di posisi bawah [7].

Penelitian yang serupa pernah dilakukan oleh Sciandra & Spera yang mempelajari ketergantungan popularitas dari musik dengan menggunakan extensi dari *Beta regression model*, termasuk *random effect*-nya. Hasil dari model ini adalah *Beta model with mixed effect* (Beta GLMM). Hasil penelitiannya mengatakan bahwa tidak semua karakteristik Spotify mempunyai *explanatory power* yang tinggi untuk jumlah *stream* yang lebih tinggi, tapi beberapa yang lain ternyata berpengaruh penting [15].

Dengan menggunakan teknik *data analytics* yaitu *data mining* peneliti akan menganalisa data dari dataset yang diambil dari *Spotify* Web API. Metode pertama yang dipakai adalah *Multiple Regression* yang merupakan pengembangan dari *Simple Linear Regression* (SLR). SLR merupakan metode statistik yang berfungsi untuk menguji sejauh mana hubungan sebab akibat antara Variabel Faktor Penyebab (X) atau *Predictor* atau *Independent Variable* terhadap Variabel Akibatnya (Y) atau *Response* atau *Dependent Variable*. *Multiple Regression* dipakai untuk memprediksi *value* berdasarkan dua atau lebih variabel *predictor*. Metode kedua yang dipakai adalah *Polynomial Regression*. Regresi polinomial merupakan regresi berganda yang dibentuk dengan menjumlahkan pengaruh variabel prediktor (X) yang dipangkatkan secara meningkat sampai orde ke-k. Metode regresi sudah pernah dipakai dalam penelitian sebelumnya, seperti contoh penelitian yang dilakukan oleh Nijkamp di tahun 2018 yang menggunakan stream counts sebagai measure popularitasnya [12].

1.2 Perumusan Masalah

Apa hubungan antara audio features dan popularitas sebuah lagu, dan juga seberapa akurat metode regresi dalam memprediksi popularitas sebuah lagu.

1.3 Tujuan Penelitian

Tujuan dari skripsi ini adalah untuk memprediksi nilai popularitas dari sebuah lagu berdasarkan audio features-nya dan menentukan tingkat akurasi dari *Multiple Regression* dan *Polynomial Regression* dalam kasus ini.

2. LANDASAN TEORI

2.1 Multiple Regression Statistical Method

Multiple linear regression adalah analisis statistik yang digunakan untuk mengetahui pengaruh beberapa variabel bebas (*independent*) terhadap variabel terikat (*dependent*) [6]. Perbedaannya dari *linear regression* yang biasa adalah MLR harus bisa meng-handle input yang banyak.

Persamaan regresi secara umum adalah:

$$y = B * x + A \quad (1)$$

Di mana:

y = variabel terikat

x = variabel bebas

A = intercept atau konstanta

B = koefisien regresi (kemiringan)

Sedangkan *multiple regression* persamaannya adalah sebagai berikut:

$$y = B\_1 * x\_1 + B\_2 * x\_2 + \ldots + B\_n * x\_n + A \quad (2)$$

Di mana:

x_1 = variabel bebas pertama

x_2 = variabel bebas kedua

x_n = variabel bebas terakhir

B_1, B_2, B_n = koefisien yang ada di setiap variabel bebas x

Masalah yang bisa muncul dalam analisa regresi adalah tentang perbedaan range data. Dalam penelitian yang dilakukan oleh da Silva dan Seixas di tahun 2017, disimpulkan bahwa menambah range data lebih efektif daripada menambah data points untuk mengurangi ketidakpastian dari kemiringan garis regresi [4]. Jadi range data yang berbeda tidak berarti regresi tidak mungkin untuk dilakukan.

2.2 Polynomial Regression Statistical Method

Polynomial Regression merupakan sebuah bentuk analisa regresi di mana hubungan antara variabel bebas dan variabel terikatnya dimodelkan di dalam orde polinomial (nth *degree polynomial*). Regresi polinomial merupakan kasus khusus dari regresi linear di mana kita bisa melakukan *fitting* persamaan polinomialnya ke data dengan hubungan *curvilinear* antara variabel terikat dan bebasnya [1].

Persamaannya adalah sebagai berikut:

$$y = b_0 + b_1x_1 + b_2x1^2 + \ldots + b_nx_1^n \quad (3)$$

Di mana:

y = variabel terikat

x = variabel bebas

b = koefisien regresi

n = orde polimonials

2.3 Feature Selection

Feature selection merupakan proses mengidentifikasi dan memilih variabel-variabel input yang paling relevan dengan target variabel [3]. Dengan menggunakan metode *Mutual Information* yang merupakan metode untuk menghitung ketergantungan antara variabel. *Mutual Information* dihitung antara 2 variabel dengan mengukur pengurangan ketidakpastian dari satu variabel jika diketahui nilai dari satu variabel yang lainnya.

2.4 Categorical Coding System

Categorical coding system mengukur data kategorikal dengan cara meng-*assign* angka numerik ke kategori tersebut, yang

mengakibatkan linear regression menjadi optimal. Hal ini disebut juga dengan CATREG [8].

Karena *categorical data* tidak begitu saja bisa dimasukkan ke dalam persamaan regression maka perlu *Coding System* untuk mengkodenya ke dalam *series of variables*. Salah satu metode yang bisa dipakai adalah *Simple Coding*.

Dalam contoh penjelasan tabel di gambar 1. (diambil dari Institute for Digital Research and Education), akan dilakukan analisa regresi. Data kategorinya mempunyai 4 level, yaitu 1 = *Hispanic*, 2 = *Asian*, 3 = *African American*, 4 = *white*. Variabel terikatnya adalah *write* yang merupakan skor *reportwriting* dari setiap ras. Dalam menggunakan metode ini dipilih dahulu 1 kategori sebagai *reference level*. Untuk kasus ini, diberi contoh dengan memilih ras *white* sebagai reference level-nya. Kemudian akan dibuat 3 variabel baru yaitu x1, x2, x3. Perlu diingat bahwa sum dari setiap variabel haruslah = 0. Akibatnya setiap level / grup ras memiliki perbandingan yang harus disesuaikan.

Untuk x1 mempunyai *coding* ¾ terhadap grup 1 (*Hispanic*), dan -¼ untuk grup lainnya.

Untuk x2 mempunyai *coding* ¾ terhadap grup 2 (*Asian*), dan -¼ untuk grup lainnya.

Untuk x3 mempunyai *coding* ¾ terhadap grup 3 (*African American*), dan -¼ untuk grup lainnya.

Level of race	New variable 1 (x1)	New variable 2 (x2)	New variable 3 (x3)
1 (Hispanic)	.75	-.25	-.25
2 (Asian)	-.25	.75	-.25
3 (African American)	-.25	-.25	.75
4 (white)	-.25	-.25	-.25

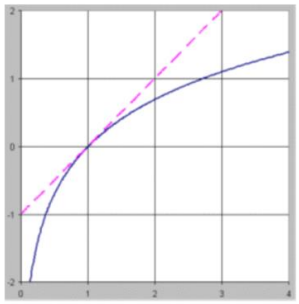
Gambar 1. Gambar Tabel Simple Regression Coding

2.5 Asumsi Regresi

Sebelum masuk ke proses regresi, perlu proses *preprocessing* untuk memastikan data sesuai dengan asumsi regresi. Ada 4 asumsi esensial yang seharusnya dipenuhi sebelum membuat model untuk prediksi [10]. Asumsi itu adalah:

1. Linearity

Merupakan hubungan linear antara variabel bebas dan terikat. Asumsi ini dapat dicek dengan menggunakan *Plot Observed Value vs Predicted Value* dan juga *Plot Residual vs Predicted*. Dalam kebanyakan kasus dunia nyata, lebih spesifik lagi dalam penelitian *audio features spotify* ini data yang muncul mustahil akan terbentuk garis lurus. Tentunya ada yang harus dilakukan oleh peneliti untuk memastikan model regresinya bisa mempunyai prediksi yang tepat (mengurangi *error*). Cara memperbaikinya adalah dengan mengaplikasikan *nonlinear transformation* terhadap variabel bebas dan atau variabel terikat. Log *transformation* merupakan salah satu cara yang efektif dalam memperbaiki pelanggaran asumsi ini. LOG dan LN akan dipakai untuk mengacu pada fungsi *natural* log. Dan simbol "≈" berarti 'kira-kira sama dengan'.



Gambar 2. Log *transformation*

Di Gambar 2. terlihat bahwa:

$$LN(1+r) \approx r \quad (4)$$

Ketika r jauh lebih kecil dari 1 dalam skala. Misalnya x naik dengan persentase kecil, seperti 5%. Ini berarti bahwa X berubah dari X menjadi X(1+r), di mana r = 0.05

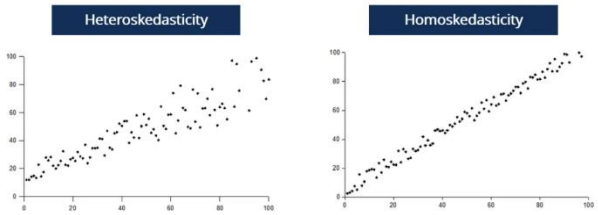
% change in X	diff-log of X
-50%	-0.693
-40%	-0.511
-30%	-0.357
-20%	-0.223
-10%	-0.105
-5%	**-0.051**
-2%	**-0.020**
0%	**0.000**
2%	**0.020**
5%	**0.049**
10%	0.095
20%	0.182
30%	0.262
40%	0.336
50%	0.405
100%	0.693

Gambar 3. Tabel Perbandingan Transformasi x dan *Diff-log*

Dalam Gambar 3. terlihat bahwa perubahan persentase dan *diff-logs* adalah sama persis di rentang +/-5% dan sangat dekat sampai di rentang +/-20%.

2. Homoscedasticity

Merupakan varians *error* yang konstan terhadap semua nilai x.



Pelanggaran terhadap asumsi ini membuat sulitnya menentukan standar deviasi yang benar dari error. *Confidence interval* bisa menjadi terlalu lebar ataupun terlalu sempit. Cara memperbaikinya bisa dengan menerapkan log transformation seperti yang dijelaskan di poin *linearity*.di poin *linearity*.

Gambar 4. Perbandingan *Homoscedasticity* dan *Heteroscedasticity*

Cara mendeteksi *Heteroscedasticity* adalah dengan menggunakan metode *Breusch Pagan.* Tes *Breusch-Pagan-Godfrey* merupakan sebuah tes terhadap eror *Heteroscedasticity* dalam regresi [17]. Nilai minimal P-Value untuk menerima hipotesa *Homoscedasticity* adalah 0.05, yang berarti jika nilai P-Value-nya lebih kurang dari 0.05 maka bisa dikatakan *Heteroscedasticity* terjadi di dalam

regresi. *Heteroscedasticity* juga bisa terlihat secara visual lewat plot seperti pada gambar 4.

3. Independence
Merupakan variabel yang independen satu dengan yang lainnya. Contohnya model secara sistematis akan melakukan *overprediction* atau *underprediction* jika variabel independen mempunyai konfigurasi tertentu. Asumsi ini dapat dicek dengan menggunakan metode *Variance Inflation Factor* (VIF). VIF merupakan suatu metode perhitungan *multicollinearity* dalam suatu set variabel *multiple regression* [14]. VIF menghitung seberapa besar nilai variance yang meningkat. Nilai VIF ini akan dilihat di setiap variabel independent (prediktor) dalam model *multiple regression*. Nilai VIF 1 berarti tidak ada korelasi sama sekali antara variabel prediktor. *Rule of thumb* secara umum jika nilai VIF 4 maka perlu investigasi lebih lanjut sedangkan nilai VIF melebihi 10 menandakan adanya *multicollinearity* yang serius.

4. Normality
Merupakan *error* yang terdistribusi secara normal. Secara teknis, asumsi ini sering tidak digunakan dalam key assumption dalam analisa regresi jika tujuannya untuk mengestimasi koefisien dan men-*generate* prediksi seperti meminimalkan *mean squared error*. Asumsi ini dapat dicek dengan menggunakan melihat nilai Omnibus. Omnibus merupakan tes untuk menghitung distribusi eror yang mengindikasikan *normality* [11]. Nilai yang diharapkan adalah mendekati 0 yang berarti *normality* terpenuhi. Cara memperbaiki adalah dengan mengaplikasikan log *transformation* seperti yang dijelaskan di poin *linearity*.

Dalam data yang diambil dari *real world* bisa dapat dipastikan perlu dilakukannya transformasi terlebih dahulu agar supaya asumsi-asumsi tersebut dapat terpenuhi. Strategi yang bisa dilakukan adalah dengan menggunakan Log Transformation.

Log Transformation merupakan metode statistikal yang sangat penting dalam pemodelan [5]. Ada 3 logaritma yang biasanya dipakai sebagai standar. Base-2 (sering dipakai dalam *computer science* dan *music theory*), Base-10 (sering dipakai dalam *engineering*), dan Natural logarithm (sering dipakai dalam matematika dan bisnis ekonomi).

2.6 Proses regresi
Proses regresi dimulai dari memilih variabel *dependent* (*outcome measure*) sesuai dengan studi kasus yang diteliti. Kemudian dipilih variabel *predictors* (*independent*) yang bertujuan untuk memaksimalkan nilai *coefficient of determination* R2 (semakin tinggi nilainya maka semakin sedikit *error* yang ada, dan itu berarti prediksinya semakin akurat) [13]. Metode regresi ini juga memungkinkan untuk memasukan banyak variabel prediktor dalam model prediksinya. Variabel-variabel prediktor yang menghasilkan prediksi paling bagus yang akan dipakai dalam model akhirnya.

Untuk menguji akurasi dari prediksi maka akan dilakukan proses *assessment* dengan menggunakan metrik-metrik evaluasi. Beberapa contoh metrik evaluasi yang bisa dipakai adalah *Mean Squared Error* (MSE), *Mean-Absolute-Error* (MAE), R^2 atau *Coefficient of Determination*, dan *Adjusted* R2.

2.7 Metrik evaluasi regresi
Mean Squared Error: Merupakan *average dari squared difference* antara *target value* dan *value* yang diprediksi oleh model regresi. Semakin kecil nilai MSE berarti estimasi menjadi semakin baik.

Mean Absolute Error: Merupakan *absolute difference* antara target value dan value yang diprediksi oleh model regresi. Paling intuitif karena hanya melihat data absolutnya. Kurang cocok buat *outlier* karena kurang sensitif.

R^2 *Error*: *Coefficient of Determination* atau R^2 : Membandingkan model dengan sebuah constant baseline dan memberi tahu seberapa lebih baik model kita. *Constant baseline* dipilih dengan mengambil *mean of the data* dan menggambarkan garis pada *mean* tersebut.

Adjusted R2: Menghitung proporsi variasi dari variabel dependent berdasarkan semua variabel *independent*. Yang membedakan R dan R2 adalah metrik R2 lebih baik digunakan dalam memilih variabel *predictors* (*independent*) karena nilai dari R2 akan naik hanya jika variabel tersebut berdampak signifikan terhadap variabel *dependent*.

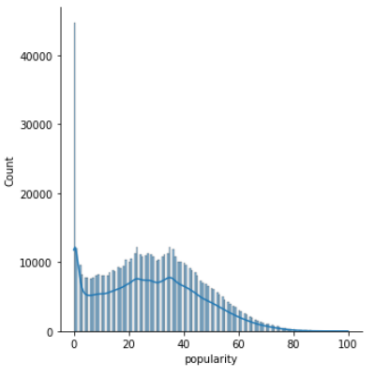
3. ANALISIS SISTEM
3.1 Analisa permasalahan
Analisis masalah merupakan investigasi terhadap persoalan yang muncul dalam penelitian *audio features Spotify* dan mengindentifikasi kemungkinan-kemungkinan solusi yang dapat digunakan untuk mengatasi masalah tersebut. Analisa metode akan lebih lanjut dijelaskan di dalam bab ini.
Data *audio features* yang diambil dari *Spotify* sangatlah beraneka ragam *range* nilainya. Ini terjadi karena data *real world* biasanya memang demikian, apalagi penelitian di bidang musik ini yang merupakan sebuah karya seni. Ini yang menjadi masalah utama dalam menjalankan prediksi popularitas lagu.

3.2 Analisa kebutuhan
Analisis kebutuhan merupakan proses penggambaran dari aktivitas yang akan diimplementasikan dalam sebuah sistem dan juga menjelaskan kebutuhan yang dibutuhkan oleh sistem agar sistem dapat berjalan sesuai dengan kebutuhan. Analisis kebutuhan ini meliputi analisis kebutuhan data dan pemodelan sistem. Sesuai dengan analisa permasalahan di poin 3.1 maka akan dijelaskan tentang kebutuhan-kebutuhan dalam penelitian ini. Data yang dibutuhkan haruslah dalam pembagian yang merata dalam nilai popularitasnya, agar hasilnya tidak menjadi bias. Jadi peneliti mengambil pembagian range 25% untuk popularitas 0-25, selanjutnya 25% untuk popularitas 26-50, selanjutnya 25% untuk popularitas 51-75, dan terakhir 25% untuk popularitas 76-100. Data yang sudah ada pun haruslah ditranformasi sesuai dengan kaidah regresi agar hasilnya menjadi tepat sesuai dengan yang seharusnya. Metode untuk melakukan transformasi (*preprocessing*) akan dijelaskan lebih lanjut di poin berikutnya.

3.3 Analisa Preprocessing
Proses *preprocessing* merupakan suatu proses yang sangat penting dalam *data analysis* agar supaya data yang dipakai menjadi tepat, efektif, dan sesuai dengan kebutuhan. Biasanya data yang dikumpulkan dalam suatu penelitian mengandung *noise* atau bahkan juga *missing*. Di sinilah salah satu peran proses *preprocessing* untuk mengatasi hal tersebut. Tapi dalam penelitian ini, data yang didapat dari Spotify sudah dalam bentuk yang sesuai dan tanpa adanya data yang *missing* atau muncul *noise*. Proses preprocessing di sini adalah untuk memastikan data yang akan diolah sudah sesuai dengan kebutuhan yaitu untuk analisa regresi. Variabel-variabel yang bertipe kategorikal harus dikodekan terlebih dahulu agar bisa dimasukkan ke dalam model regresi dan juga dapat dilakukan transformasi terhadap data untuk mendapatkan hasil yang lebih baik. Selanjutnya, ada *requirement* mendasar yang seharusnya dipenuhi dalam analisa regresi. *Requirement* itu adalah asumsi regresi.

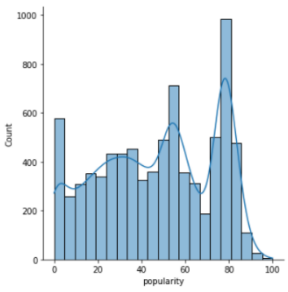
3.3.1 Dataset Preparation
Sebelum data diolah dengan proses lebih lanjut, maka data yang sudah didapatkan perlu dipersiapkan terlebih dahulu. Data lagu

yang diambil dari data *Spotify* mempunyai *range* popularitas yang sangat bervariasi dan tidak seimbang.



Gambar 5. *Range* popularitas data sebelum pembagian

Terlihat pada gambar 5. yang menunjukkan bahwa lagu dengan popularitas 0 mempunyai jumlah yang sangat banyak. Dari total 586.672 baris lagu, ternyata lebih dari 40.000 baris lagu mempunyai nilai popularitas 0. Peneliti melakukan pembagian terlebih dahulu untuk memastikan *range* popularitas lagunya menjadi lebih seimbang untuk kebutuhan analisa. Jumlah baris dan juga pembagian persentase popularitasnya dipillih sesuai dengan keadaan dataset. Dataset lagu tersebut mempunyai kecenderungan condong ke nilai popularitas di bawah 40.



Gambar 6. *Range* popularitas data setelah pembagian

Dengan menerapkan pembagian popularitas dengan persentase masing-masing 25% untuk *range* 0-25, 26-50, 51-75, 76-100 maka didapatkanlah hasil seperti pada gambar 6. Terlihat dengan data yang sudah dibagi ini, mempunyai distribusi popularitas yang lebih seimbang.

3.3.2 Simple Coding

Variabel kategorikal membutuhkan perlakuan khusus dalam analisa regresi karena tidak bisa begitu saja dimasukkan ke dalam persamaan regresi. Variabel kategorikal ini harus dikodekan ke dalam *series of variables*. Berikut merupakan contoh cara kerja metode ini dalam bentuk variabel nada dasar lagu (*music key*).

Dalam metode ini, variabel input dari data disebut sebagai *reference group*. Misalnya input grup 0 (kunci C) sebagai *reference*. Variabel x1 merupakan perbandingan grup 0 dan grup 1, variabel x2 merupakan perbandingan grup 0 dan grup 2, variabel x3 merupakan perbandingan grup 0 dan grup 3, dan seterusanya. Variabel x0 *coding*-nya adalah ¾ untuk grup 0 dan -1/4 untuk grup lainnya, variabel x1 *coding*-nya adalah ¾ untuk grup 1 dan -1/4 untuk grup lainnya, variabel x2 *coding*-nya adalah ¾ untuk grup 2 dan -1/4 untuk grup lainnya, dan seterusnya. Perlu diperhatikan bahwa setiap variabel harus berjumlah 0 jika dijumlahkan.

Tabel 1. Pengkodean Variabel Kategorikal

Level of music key	X0	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
0 (C)	**11/12**	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12
1 (C#)	-1/12	**11/12**	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12
2 (D)	-1/12	-1/12	**11/12**	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12
3 (D#)	-1/12	-1/12	-1/12	**11/12**	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12
4 (E)	-1/12	-1/12	-1/12	-1/12	**11/12**	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12
5 (F)	-1/12	-1/12	-1/12	-1/12	-1/12	**11/12**	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12
6 (F#)	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	**11/12**	-1/12	-1/12	-1/12	-1/12	-1/12
7 (G)	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	**11/12**	-1/12	-1/12	-1/12	-1/12
8 (G#)	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	**11/12**	-1/12	-1/12	-1/12
9 (A)	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	**11/12**	-1/12	-1/12
10 (A#)	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	**11/12**	-1/12
11 (B)	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	-1/12	**11/12**

Variabel-variabel baru ini siap digunakan untuk dimasukkan ke dalam model regresi ataupun digunakan untuk proses preprocessing lebih lanjut.

3.3.3 Log Transformation

Untuk mentransformasi variabel-variabel prediktor yang ada di dalam dataset, digunakan metode logaritma dengan basis 10. Berikut ini contoh perhitungan *log transformation*:

- Contoh variabel *Speechiness* mempunyai data 0.0838, 0.0522, 0.0768.
  Dengan $\log_{10}$ maka didapatkan hasil:
  0.0838 -> -1.076756
  0.0522 -> -1.282329
  0.0768 -> -1.114639
- Contoh variabel *Speechiness* yang dikalikan 100 mempunyai data 8.38, 5.22, 7.68.
  Dengan $\log_{10}$ maka didapatkan hasil:
  0.0838 -> 0.923244
  0.0522 -> 0.717671
  0.0768 -> 0.885361

Log transformation tersebut dilakukan terhadap variabel *danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration_ms*. Beberapa pengujian juga dilakukan hanya dengan variabel tertentu, misalnya *acousticness* atau *speechiness* saja untuk melihat outputnya terhadap metriks uji coba.

3.4 Analisa Metode

3.4.1 Multiple Regression

Metode multiple regression dilakukan dengan memilih variabel-variabel bebas dan 1 variabel terikat.

Variabel terikat (y) merupakan variabel yang hasilnya didapat dari model persamaan variabel bebas (x), koefisien (B), dan juga konstanta (A).

3.4.1.1 Estimasi dari Parameter Model

•Estimasi dari nilai koefisien B merupakan nilai yang meminimalisir *sum* dari *squared error*.

•Huruf b nantinya dipakai untuk merepresentasikan estimasi sampel dari koefisien B. Jadi b0 merupakan estimasi sampel dari B0, b1 merupakan estimasi sampel dari B1, dan seterusnya.

•MSE mengestimasi σ2, *variance of the errors*.

•Dalam kasus yang mempunyai 2 prediktor, persamaan regresinya menghasilkan sebuah *plane*, sedangkan jika lebih dari 2 prediktor akan menghasilkan sebuah *hyperplane*.

3.4.1.2 Menginterpretasi Parameter Model

•Setiap koefisien B merepresentasikan perubahan di mean response, E(y), setiap unit akan naik jika semua prediktor yang lain nilainya konstan. Misalnya, B1 merepresentasikan *mean response*, setiap unit akan naik di x1 jika x2, x3, …, xn konstan.

•B0 atau yang disebut intercept merepresentasikan *mean response*, jika semua prediktor x1, x2, … xn nilainya 0.

3.4.1.3 Fitted Values dan Residuals

•Nilai fitted atau prediksi dihitung sebagai $\hat{Y}i=b0+b1xi,1+b2xi,2+\ldots+bnxi,n$,, di mana nilai b didapat dari *software* statistika atau *library*, dan nilai x ditetentukan oleh peneliti.

•Nilai residual dihitung sebagai $ei=Yi-\hat{Y}$, yang merupakan perbedaan dari nilai sebenarnya dan nilai prediksi.

•Plot dari residual versus nilai prediksi secara ideal sebaiknya menyerupai garis *horizontal* yang *random* (*horizontal random band*)

3.4.2 Polynomial Regression

Metode *polynomial regression* merupakan metode yang dipakai untuk mengatasi adanya hubungan variabel prediktor yang *nonlinear*.

3.4.2.1 Estimasi Regresi Polynomial

•Fitted model akan lebih reliable jika ukuran sampel n yang lebih besar

•Jangan meramalkan melebihi batas nilai yang bisa diobservasi, karena persamaan polynomial mempunyai *curve* yang sangat tegas. Jika meramalkan lebih dari itu akan menghasilkan hasil yang tidak ada artinya.

3.4.2.2 Model Building Strategy

•*Forward Selection*. Fit model dengan order yang meningkat sampai tes yang paling tinggi sudah tidak signifikan lagi.

•*Backward Elimination*. Fit model dari order paling besar, kemudian menghapus satu per satu ordernya dari yang paling tinggi sampai mendapatkan hasil order tertinggi yang paling signifikan

4. PENGUJIAN SISTEM

Model-model yang ada akan diuji akurasinya dengan melihat nilai metrics dari setiap pemilihan variabel dan langkah *preprocessing* yang dilakukan. Pengujian dijalankan pada *device* HP Omen 15 dengan spesifikasi prosesor 3.0 Ghz AMD Ryzen 5 4600H dan memori RAM 16GB DDR4.

Data yang digunakan untuk pengujian merupakan data yang berasal dari Spotify. Total data berjumlah 565.655. Tetapi data yang akan diuji akan dipilih sesuai dengan kebutuhan dan efektifitas pemrosesan.

Proses *features selection* dijalankan sebanyak 5 kali untuk mendapatkan hasil yang lebih akurat. Setelah dijalankan sebanyak 5 kali, maka didapatkan nilai rata-rata dengan detail seperti terlihat pada tabel 2.

Dari hasil tabel 2. tersebut terlihat bahwa hasilnya konsisten dalam 5 kali *running* proses *feature selection*. Peneliti mengambil 8 variabel berdasarkan proses ini dengan melihat nilai terendahnya yaitu 0.0208 untuk variabel danceability. Variabel yang punya nilai di bawah 0.001 tidak diambil dalam beberapa uji coba.

Tabel 2. Feature Selection

No	Audio Features	1	2	3	4	5	Average
0	**Acousticness**	**0.127944**	**0.126211**	**0.126105**	**0.125959**	**0.126862**	**0.126616**
1	**Danceability**	**0.020314**	**0.022655**	**0.021298**	**0.019233**	**0.020961**	**0.020892**
2	Mode	0.000785	0.000000	0.000248	0	0.001342	**0.000475**
3	**Energy**	**0.066354**	**0.068378**	**0.068031**	**0.068524**	**0.066084**	**0.067474**
4	**Instrumentalness**	**0.049800**	**0.045054**	**0.048393**	**0.047974**	**0.046547**	**0.047554**
5	Liveness	0.008905	0.008727	0.010085	0.007566	0.009047	**0.008866**
6	**Loudness**	**0.078036**	**0.078132**	**0.077608**	**0.078547**	**0.077921**	**0.078049**
7	**Speechiness**	**0.037506**	**0.037135**	**0.035712**	**0.038863**	**0.035896**	**0.037022**
8	**Tempo**	**0.045362**	**0.045575**	**0.045267**	**0.045526**	**0.045544**	**0.045455**
9	**Duration_ms**	**0.066566**	**0.066637**	**0.066851**	**0.066675**	**0.066666**	**0.066679**
10	Valence	0.007132	0.008075	0.009339	0.007721	0.008211	**0.008096**
11	Key	0.006044	0.006993	0.007038	0.008252	0.004457	**0.006557**

Kedua metode (model terbaik dari masing-masing metode) akan dibandingkan dengan melihat rata-rata metrik dari hasil ujicoba yang sudah dilakukan.

Tabel 3. Tabel Perbandingan Rata-Rata Hasil Metriks 2 Metode

No	Metrics	Linear Regression	Polynomial Regression
1	Average R2 Train	0.23783	0.34185
2	Average MSE	442.31700	409.76242
3	Average MAE	17.38129	16.47367
4	Average R2	0.23614	0.31496
5	Average Adjusted R2	0.23536	0.25880

Berdasarkan hasil rata-rata dari pengujian yang sudah dilakukan, metode Polynomial Regression mempunyai nilai yang lebih baik di semua metriks pengujian. Hasil ini konsisten di hampir semua *testing* dan *subtesting*.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan implementasi dan pengujian dari Analisa Audio Features dengan Membandingkan Metode Multiple Regression dan Polynomial Regression untuk Memprediksi Popularitas Lagu yang dilakukan maka dapat disimpulkan bahwa:

- Terdapat hubungan yang relatif sedang antara audio features dan popularitas yang terlihat lewat variance yang mampu dijelaskan oleh model dengan nilai 0.4348 pada ujicoba *Polynomial Regression* dengan menggunakan *Simple Coding*.
- Pada pengujian dengan menggunakan *Simple Coding* menghasilkan nilai R2 dan *Adjusted* R2 yang cenderung lebih tinggi.
- Hampir mustahil untuk bisa membangun model yang memenuhi semua asumsi klasikal regresi dikarenakan data lagu / musik merupakan sebuah karya seni di mana bersifat lebih abstrak dan subjektif.

- Berdasarkan hasil pengujian, terlihat bahwa model *Polynomial Regression* mempunyai akurasi yang lebih baik di semua metriks pengujian.

5.2 Saran

Berikut merupakan beberapa saran yang dapat diberikan oleh penulis untuk pengembangan aplikasi dan penelitian ini ke depannya.

- Dapat menganalisa dengan membagi *dataset* berdasarkan *genre* lagunya dan juga *region* lagunya. Dengan demikian maka analisa audio features untuk memprediksi popularitas menjadi lebih jelas karena setiap *genre* pasti memiliki karakteristik audionya dan setiap daerah mempunyai seleranya yang unik.
- Dapat menganalisa lebih dalam lagi hubungan antara asumsi klasikal regresi dengan data *audio features Spotify*.
- Dapat mengembangkan aplikasi yang terhubung secara komplit antara program analisa dan program *user interface*-nya, di mana user bisa melakukan analisa dengan mudah lewat tampilan UI.
- Memperdalam dan memperbanyak analisa transformasi data dengan menggunakan transformasi log agar supaya menghasilkan data yang lebih baik untuk prediksi.

6. DAFTAR PUSTAKA

[1] Abhigyan. 2020. Understanding Polynomial Regression. Retrieved February 23, 2021. URI = https://medium.com/analytics-vidhya/understanding-polynomial-regression-5ac25b970e18

[2] Armstrong, M. 2020. The world's most popular music streaming services. Retrieved January 6, 2021. URI = https://www.statista.com/chart/20826/music-streaming-services-with-most-subscribers-global-fipp/

[3] Brownlee, J. 2020. How to Perform Feature Selection for Regression Data. URI = https://machinelearningmastery.com/feature-selection-for-regression-data/

[4] da Silva, M. A. S., & Seixas, T. M. 2017. The Role of Data Range in Linear Regression. *The Physics Teacher, 55(6), 371–372*. DOI = https://doi.org/10.1119/1.4999736

[5] Duke.edu. 2021. Testing the assumptions of linear regression. URI = http://people.duke.edu/~rnau/testing.htm

[6] Grant, P. 2019. Understanding Multiple Regression. URI = https://towardsdatascience.com/understanding-multiple-regression-249b16bde83e

[7] Herremans, D., Martens, D., & Sörensen, K. 2014. Dance Hit Song Prediction. *Journal of New Music Research*, *43*(3), 291–302. DOI = https://doi.org/10.1080/09298215.2014.881888

[8] Institute for Digital Research and Education (n.d.). Coding Systems for Categorical Variables in Regression Analysis. URI = https://stats.idre.ucla.edu/spss/faq/coding-systems-for-categorical-variables-in-regression-analysis-2/#SIMPLE%20EFFECT%20CODING

[9] International Federation of the Phonographic Industry. 2019. annual report. URI = https://www.ifpi.org/our-industry/industry-data/

[10] JMP.com (n.d.). Regression Model Assumptions. URI = https://www.jmp.com/en_us/statistics-knowledge-portal/what-is-regression/simple-linear-regression-assumptions.html

[11] McCarty, K. 2018. Interpreting Results from Linear Regression – Is the data appropriate? Accelebrate.com; Accelebrate. Retrieved February 23, 2021. URI = https://www.accelebrate.com/blog/interpreting-results-from-linear-regression-is-the-data-appropriate

[12] Nijkamp, R. 2018. Prediction of product success: explaining song popularity by audio features from Spotify data. URI = https://essay.utwente.nl/75422/1/NIJKAMP_BA_IBA.pdf

[13] Palmer, P. B., & O'Connell, D. G. 2009. Regression analysis for prediction: understanding the process. *Cardiopulmonary Physical Therapy Journal, 20(3), 23–26*. URI = https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2845248/

[14] PennState: Eberly College Of Science. 2018. Detecting Multicollinearity Using Variance Inflation Factors | STAT 462. Retrieved February 21, 2021. URI = https://online.stat.psu.edu/stat462/node/180/

[15] Sciandra, M., Spera, I. C. 2019. A model based approach to Spotify data analysis: a Beta GLMM. *Journal of Applied Statistics*. DOI = https://doi.org/10.2139/ssrn.3557124

[16] Spotify for developers (n.d.). api documentation. URI = https://developer.spotify.com/documentation/web-api/reference/tracks/

[17] Statistics How To. 2016. Breusch-Pagan-Godfrey Test: Definition. URI = https://www.statisticshowto.com/breusch-pagan-godfrey-test/