

# Prediksi Peringkat Mingguan Lagu Pada Spotify Amerika Serikat Menggunakan Multiple Charts Dataset Dengan Berbagai Metode

Christianto Aryanto, Henry Novianus Palit, Andre Gunawan  
Program Studi Informatika Fakultas Teknologi Industri Universitas Kristen Petra  
Jl. Siwalankerto 121 – 131 Surabaya 60236  
Telp. (031) – 2983455, Fax. (031) – 8417658

E-Mail: christianto.aryanto@gmail.com, hnpalit@petra.ac.id, andre.gunawan@petra.ac.id

## ABSTRAK

Mayoritas pendapatan industri musik pada tahun 2020 berasal dari *streaming music* dengan presentase 62.1%. Hal ini membuat banyak pihak industri musik yang berlomba-lomba untuk bisa mendapatkan lagu hit, terutama di tangga lagu/chart Spotify Amerika Serikat. Akan tetapi, hal ini sulit dilakukan karena zaman sekarang performa suatu lagu tidak bisa hanya ditentukan dari kualitas sebuah lagu, melainkan ditentukan oleh performa lagu diberbagai *music chart*. Oleh karena itu, akan dilakukan penelitian pada bidang *hit song science* untuk memprediksi peringkat mingguan lagu pada Spotify Amerika Serikat dengan menggunakan dataset chart dari Spotify, Shazam, Airplay, dan TikTok.

Penelitian ini akan menggunakan metode *multipler linear regression*, *polynomial regression*, *gradient boosting tree*, dan *random forest* dalam pembuatan model dan akan dibandingkan performanya menggunakan *adjusted r-squared* dan *mean absolute error* (MAE) sebagai evaluation metric. Model terbaik dihasilkan oleh metode *random forest* dengan nilai *adjusted r-squared* dan MAE sebesar 93.133% dan 11.687. Penggunaan dataset atribut lagu memberikan dampak buruk pada kinerja model. Disisi lain, Penggunaan dataset Shazam chart memberikan dampak yang baik pada kinerja model. Sedangkan, penggunaan dataset Airplay dan TikTok chart tidak memberikan dampak yang jelas. Akan tetapi, kedua chart ini terbukti memiliki relasi yang lemah terhadap model. Secara keseluruhan, penggunaan dataset Spotify, Shazam, Airplay, dan TikTok chart merupakan kombinasi dataset yang menghasilkan model terbaik pada penelitian ini.

**Kata Kunci:** Spotify, *hit song science*, *ranking prediction*, *multiple linear regression*, *polynomial regression*, *gradient boosting tree*, *random forest*

## ABSTRACT

*In 2020, the majority of the music industry's revenue, 62.1%, came from streaming music. As a result, many music business parties are striving for a hit song, particularly on Spotify US chart. However, this is difficult to achieve because nowadays, a song's performance is determined by its performance on various music charts, not by its quality. Due to that, a study in the field of hit song science will be conducted to forecast weekly song ranking on Spotify US using data from Spotify, Shazam, Airplay, and TikTok charts.*

*Multipler linear regression, polynomial regression, gradient boosting tree, and random forest are the methods used in this study to create models, and each model will be compared using adjusted r-squared and mean absolute error (MAE) as evaluation metrics. Random forest produced the best model, with adjusted r-squared*

*and MAE values of 93.133% and 11.687, respectively. The usage of music attribute had a negative impact on model performance. Shazam chart, on the other hand, has been shown to have a positive impact on model performance. Meanwhile, neither the Airplay nor the TikTok charts have a definite positive or negative impact. However, both have been shown to have a very weak relation with model performance. Overall, the dataset combination of Spotify, Shazam, Airplay, and TikTok chart produced the best model in this study.*

**Keywords:** Spotify, *hit song science*, *ranking prediction*, *multiple linear regression*, *polynomial regression*, *gradient boosting tree*, *random forest*

## 1. PENDAHULUAN

Pendapatan industri musik mengalami peningkatan setiap tahunnya, dengan Amerika Serikat sebagai *music market* terbesar di dunia. Pada tahun 2020, pendapatan *streaming music* telah mencapai 62.1% dari total pendapatan industri musik global [8]. Salah satu *streaming platform* terbesar saat ini ialah Spotify, dengan total pengguna bulanan aktif melebihi 340 juta. [17]. Hal ini menjadikan Spotify, terutama *chart* (papan peringkat) musik pada Spotify United States, sebagai salah satu determinan terbesar akan sukses tidaknya sebuah lagu. Saat ini, sebuah lagu hit tidak bisa didapatkan hanya dengan kualitas lagu saja, melainkan dengan faktor penentu lainnya juga, seperti popularitas artis dan performa lagu diberbagai *music chart*. Hal ini membuat pemilihan lagu serta strategi promo sulit untuk ditentukan. Kesalahan dalam melakukan promosi dapat menyebabkan kerugian yang besar bagi pihak yang terlibat.

Dengan adanya prediksi peringkat lagu, Record Label dan artis pun bisa mendapat pandangan mengenai performa dari suatu lagu kedepannya dan dapat memilih lagu serta jenis promosi (promosi yang berfokus pada Radio, Spotify, atau TikTok) yang tepat. [5, 10, 1]. Oleh karena itu, akan dilakukan penelitian dibidang *hit song science* untuk membuat model prediksi peringkat mingguan lagu pada Spotify US menggunakan 4 *dataset chart* yang berbeda (Spotify, Shazam, Airplay, & TikTok) serta *dataset* tambahan lainnya. Penelitian ini akan menggunakan 4 metode yang nantinya akan dibandingkan performanya. Karena *task* dari penelitian ini adalah *regression*, maka akan digunakan metode yang cocok dan sudah umum untuk digunakan dalam model *regression*, yaitu *multiple linear regression*, *polynomial regression*, *gradient boosting tree*, dan *random forest*.

## 2. PENELITIAN SEBELUMNYA

Sebuah penelitian di bidang HSS mencoba untuk mengklasifikasikan sebuah lagu hit atau tidak. Dengan menggunakan *dataset Billboard chart*, *Million Songs Dataset* (MSD), dan atribut lagu dari Spotify API. Model terbaik yang

dihasilkan setelah membandingkan beberapa metode adalah model *neural network* dengan akurasi 77% [5]. Klasifikasi serupa dilakukan oleh peneliti lain menggunakan *dataset Top 50 dan Viral 50* dari *Spotify chart* serta atribut lagu dari *Spotify API*. Model terbaik yang dihasilkan setelah melakukan perbandingan metode adalah model *SVM* dengan akurasi sebesar 91% [1]. Meskipun akurasinya tinggi, penelitian ini belum bisa memberikan model yang *credible*. Hal ini disebabkan karena sebuah lagu akan dianggap *hit* jika lagu tersebut masuk pada *Top 50*. Akan tetapi *dataset* yang digunakan hanya diambil dari *Top 50 dan Viral 50* saja yang mengakibatkan *imbalanced dataset* karena lebih dari 50% lagu di *dataset* dianggap sebagai lagu *hit*. Beberapa penelitian dibidang HSS lain juga melakukan penelitian terhadap *trend* berdasarkan *genre* atau lirik lagu. Akan tetapi, pendekatan ini terbukti tidak menghasilkan hasil yang baik [15]. Berdasarkan penelitian terdahulu, penelitian ini akan melakukan pendekatan yang berbeda dengan melakukan prediksi *rank* lagu pada *Spotify US* menggunakan berbagai *dataset chart*, seperti *Spotify, Shazam, Airplay, dan TikTok*, serta akan dibantu dengan *dataset* atribut lagu dan popularitas artis.

### 3. DATASET

Terdapat 2 jenis *dataset* yang dikumpulkan, yaitu *dataset chart* dan *dataset song attribute*. *Dataset chart* sendiri terbagi lagi menjadi 4, yaitu *dataset chart* dari *Spotify, Shazam, Airplay, dan TikTok*. Setiap *dataset chart* akan dikumpulkan terlebih dahulu secara bersamaan. Setelah itu, *dataset song attribute* akan dikumpulkan sesuai dengan lagu-lagu yang berada pada *Spotify chart* yang terkumpul minggu tersebut. Pengumpulan *dataset* akan dilakukan setiap minggunya sesuai dengan minggu *chart* tersebut dirilis karena terdapat beberapa data yang akan tidak valid jika dikumpulkan setelah tanggal rilis *chart*, seperti *peak* dan *weeks on chart* pada *dataset Spotify chart*. Pengumpulan data akan berlangsung selama periode 13 minggu (6 Agustus 2021 sampai 29 Oktober 2021). Pemilihan tanggal tersebut dilakukan untuk menghindari adanya bisa dari *event* tertentu, seperti *event natal* (November dan Desember) serta *event musim panas* (Juni dan Juli). *Dataset chart* akan dikumpulkan menggunakan *API* dari *Chartmetric*. Sedangkan *dataset song attribute* akan dikumpulkan menggunakan *API* dari *Spotify*. *Dataset* akan berjumlah 13 file *csv* pada setiap jenis *dataset chart* (total 52 file *csv*) dimana setiap file *csv* memiliki jumlah data 200 baris. Setelah penggabungan antar semua *dataset chart*, *dataset* akhir akan berjumlah 2600 baris data.

## 4. METODE

### 4.1 Linear Regression

*Regression analysis* adalah sebuah teknik yang umum digunakan dibidang statistik, programming, dan bidang lainnya. *Regression analysis* berfokus pada pemahaman hubungan antara satu variabel dengan variabel lain sehingga dapat terlihat pola relasi antara variabel-variabel tersebut [12]. *Linear regression* merupakan salah satu jenis dari *regression analysis* yang menjelaskan relasi *linear* antar variabel dependen dan independen. *Multiple linear regression* adalah *linear regression* yang memiliki variabel independen lebih dari satu [14]. Selain itu, terdapat juga *polynomial regression* yang digunakan untuk data dengan pola *polynomial equation* dan memiliki relasi *curvilinear* (lurus-lengkung) antar variabelnya. [13].

### 4.2 Gradient Boosting Tree

*Gradient boosting tree* adalah sebuah algoritma *ensemble* yang menggunakan *decision tree* sebagai *base learner*. Metode ini akan menggunakan *tree* biasa yang memiliki akurasi yang tidak terlalu bagus pada awal pembuatan model. Selanjutnya, residual atau nilai

residu dari *tree* sebelumnya akan digunakan pada pembuatan *tree* berikutnya. Hal ini bertujuan untuk meminimalisir nilai residu dari *tree* sebelumnya sehingga *tree* baru yang dibuat dapat memiliki akurasi yang lebih optimal. Proses ini akan diiterasikan secara terus menerus sesuai dengan maksimal kedalaman *tree* yang sudah ditentukan sampai model menjadi optimal [2].

### 4.3 Random Forest

*Random forest* adalah sebuah metode *machine learning* yang merupakan sebuah kelompok atas berbagai *tree*. Mekanisme metode ini dimulai dari pembuatan *tree* dalam jumlah besar. Pembuatan masing-masing *tree* akan menggunakan *training data* yang akan diambil secara acak menggunakan *random features selection*. Setelah itu, setiap *tree* akan *vote* untuk *class* yang paling *popular*. *Class* dengan *vote* terbanyak akan dipilih sebagai *final output* dari metode *random forest* ini [2].

### 4.4 Adjusted R-Squared

*R-squared* adalah sebuah pengukuran statistik yang digunakan untuk menjelaskan proporsi varians pada sebuah variabel dependen menggunakan variabel independen pada sebuah model regresi. Dengan menggunakan *r-squared* kita bisa mengetahui seberapa kuat sebuah variabel independen dapat menjelaskan atau memprediksi sebuah variabel dependen [9]. *R-squared* memiliki nilai antara 0% sampai 100%, dengan 100% sebagai hasil terbaik. Pada *r-squared*, penambahan variabel independen akan selalu menghasilkan penambahan nilai pada hasil *r-squared*, meskipun variabel independen tersebut tidak memiliki korelasi dengan variabel dependen. Masalah ini menyebabkan banyak pengujian model menjadi *overfitting*. Oleh karena itu, *r-squared* kemudian dimodifikasi agar mampu mengkalkulasi peran setiap variabel independen pada sebuah model. *Evaluation metric* ini kemudian disebut sebagai *adjusted r-squared* [11].

### 4.5 Mean Absolute Error (MAE)

*MAE (Mean Absolute Error)* adalah sebuah *evaluation metric* model regresi yang menghitung rata-rata dari *error* (selisih antara nilai prediksi dan nilai aktual) yang sudah dimutlakan [18]. *MAE* ini bertujuan untuk mengetahui seberapa jauh penyimpangan yang dihasilkan model prediksi [3]. Semakin kecil nilai *MAE* berartikan semakin bagus sebuah model dikarenakan penyimpangan *error* yang sedikit. Jika dibandingkan dengan *r-squared*, peneliti-peneliti seringkali tidak menggunakan *MAE* karena memberikan hasil yang belum *scaled/terskala* (sulit diukur tingkat kebagusannya dikarenakan tidak adanya batasan atau *benchmark*). Kelebihan dari *MAE* sendiri adalah *MAE* mampu menangkap *error* secara jelas yang seringkali sulit ditangkap oleh *evaluation metric* model regresi yang sudah *scaled*.

### 4.6 Chartmetric API

*Chartmetric* adalah sebuah *all-in-one platform* yang menyediakan informasi terkait *music chart* dan *social data* bagi pihak industri musik [4]. Terdapat juga *API* dari *Chartmetric* yang memungkinkan *developer* untuk mengakses data milik *Chartmetric* secara bebas. Data yang dapat diambil adalah semua data milik *Chartmetric* termasuk data *music chart* dari setiap *platform music*, atribut dari sebuah lagu, data *social* artis, dan sebagainya. *Developer* bisa menggunakan berbagai Bahasa pemrograman dalam mengaksesnya, seperti *Python, Javascript, dan R*.

### 4.7 Spotify API

*Spotify API* adalah sebuah platform yang disediakan *Spotify* agar *developer* bisa mengakses data musik yang dimiliki oleh *Spotify* secara bebas. *Developer* bisa menggunakan berbagai Bahasa

pemrograman dalam mengaksesnya, seperti Python, Javascript, dan R. Di API ini, bermacam data bisa diekstrak, mulai dari lagu, *album*, *streams*, *charts*, *playlist*, popularitas artis, dan bahkan atribut musik dari setiap lagu yang terdaftar di Spotify. Musik atribut sendiri terdiri dari *danceability*, *acousticness*, *energy*, *instrumentalness*, *liveness*, *loudness*, *speechiness*, *tempo*, *valence*, *key*, *duration*, dan *mode* [16].

#### 4.8 Data Preprocessing

Data preprocessing dimulai dengan memuat 4 dataset chart yang telah dikumpulkan. Setiap jenis dataset chart akan memiliki 13 file csv dari pengumpulan data selama periode 13 minggu. Variabel "Date" lalu ditambahkan disetiap dataset dengan tanggal chart tersebut rilis. Kemudian, 13 dataset kemudian digabungkan menjadi satu, hal ini dilakukan kepada keempat dataset chart.

Setelah itu, penambahan variabel "W-1\_Rank" dilakukan pada keempat dataset dan penambahan variabel "W+1\_Rank" dilakukan hanya pada dataset Spotify chart saja. Variabel "W-1\_Rank" adalah rank dari lagu tersebut 1 minggu sebelumnya. Sedangkan, Variabel "W+1\_Rank" adalah rank dari lagu tersebut 1 minggu berikutnya. Penambahan velocity pun dilakukan dengan perhitungan menggunakan variabel rank minggu tersebut dengan variabel "W-1\_Rank". Ketiga dataset chart kemudian diintegrasikan menjadi 1 dataset dengan dataset chart Spotify sebagai dataset utama. Integrasi dataset akan dilakukan menggunakan join-left dengan variabel "Chartmetric ID" sebagai key. Akan terdapat missing values dari penambahan variabel pada kolom rank dan velocity. Missing values pun ditangani dengan pengisian value bernilai 201 dan perhitungan ulang velocity menggunakan value rank yang baru saja diisi.

Setelah dataset utama terbentuk, dapat dilakukan penambahan variabel "Song Age" pada 2600 baris data dari dataset utama. Kemudian dilakukan data transformation, kolom yang akan ditransformasi adalah "Artist Popularity", "Duration", "Tempo", dan "Loudness". Transformasi yang dilakukan adalah normalisasi min-max dengan batasan bawah dan atas berdasarkan nilai mean, min, max, dan kuartil masing-masing kolom. Setelah itu, kolom "Weeks on Chart" dan "Song Age" di discretization dengan cara mengubah masing-masing kolom dari data numerical menjadi data ordinal categorical. Pembagian kategori akan didasari dari nilai kuartil masing-masing kolom. Terakhir, akan dilakukan data reduction untuk membuang kolom-kolom yang tidak dibutuhkan dalam implementasi model.

### 5. PENGUJIAN

Untuk memudahkan penyampaian, pada Bab ini penyampaian metode dan evaluation metric hanya akan menggunakan singkatannya saja. Untuk metode, multiple linear regression, polynomial regression, gradient boosting tree, dan random forest akan disingkat dengan MLR, PR, GBT, dan RF secara berturut-turut. Untuk evaluation metric, nilai adjusted r-squared dan mean absolute error akan disingkat dengan A-R2 dan MAE.

#### 5.1 Pengujian Model Secara Umum

Pembuatan model dilakukan pertama kali secara umum. Model secara umum yang dimaksud ialah pembuatan model sebelum dilakukannya splitting ratio selection, features selection, dan parameter tuning. Splitting ratio yang digunakan adalah 6:4. Features yang digunakan adalah semua features yang ada. Parameter yang digunakan adalah default parameter dari setiap regressor yang digunakan.

Hasil evaluation metric dari model ini dapat dilihat pada Tabel 1. Terlihat bahwa secara keseluruhan, pengujian model secara umum ini memberikan hasil rata-rata A-R2 sebesar 88.590% dan MAE sebesar 14.557. Metode dengan hasil A-R2 dan MAE terbaik adalah GBT dengan hasil 90.861% dan 13.191 secara berturut-turut.

Tabel 1. Hasil Pengujian Secara Umum

	A-R2	MAE
MLR	86.022%	16.619
PR	86.649%	15.109
GBT	<b>90.861%</b>	<b>13.191</b>
RF	90.829%	13.309
AVG	88.590%	14.557

#### 5.2 Pengujian Splitting Ratio Selection

Pada tahap ini, akan dilakukan splitting ratio selection dengan tujuan untuk mencari tahu splitting ratio mana yang paling baik untuk digunakan. Pada pengujian ini akan digunakan 3 jenis splitting ratio, yaitu 6:4, 7:3, dan 8:2. Features yang digunakan adalah semua features yang ada. Parameter yang digunakan adalah default parameter dari setiap regressor yang digunakan.

Hasil evaluation metric dari pengujian ini dapat dilihat pada Tabel 2. Splitting ratio dengan perbandingan 8:2 memiliki hasil rata-rata terbaik dengan A-R2 sebesar 88.980% dan MAE sebesar 13.801. Terlihat juga bahwa jenis splitting ratio ini memiliki 7 dari 8 evaluation metric terbaik. Hanya metric A-R2 dari metode PR dengan splitting ratio 6:4 yang memiliki nilai lebih baik dari pada splitting ratio 8:2 dengan selisih kurang dari 1%. Dari hasil ini dapat disimpulkan bahwa perbandingan 8:2 adalah splitting ratio dengan hasil paling optimal. Jenis splitting ratio ini yang akan seterusnya digunakan dalam pengujian model pada tahap pengujian berikutnya.

Tabel 2. Hasil Pengujian Splitting Ratio Selection

	6:4		7:3		8:2	
	A-R2	MAE	A-R2	MAE	A-R2	MAE
MLR	86.02%	16.62	86.60%	16.39	<b>86.98%</b>	<b>16.09</b>
PR	<b>86.65%</b>	15.11	86.34%	14.70	85.89%	<b>14.07</b>
GBT	90.86%	13.19	90.49%	13.04	<b>91.31%</b>	<b>12.53</b>
RF	90.83%	13.31	91.00%	12.94	<b>91.74%</b>	<b>12.52</b>
AVG	88.59%	14.56	88.61%	14.26	88.98%	13.80

#### 5.3 Pengujian Features Selection

Pada tahap ini, akan dilakukan features selection dengan tujuan untuk mencari tahu kombinasi variabel independen mana yang paling baik untuk digunakan. Splitting ratio yang digunakan adalah 8:2. Parameter yang digunakan adalah default parameter dari setiap regressor yang digunakan. Total variabel independen atau features yang ada pada dataset yang telah di preprocessing berjumlah 18. Apabila semua variabel independen tersebut dikombinasikan, maka akan terdapat 524,288 kombinasi variabel yang perlu dianalisa. Oleh karena itu, akan dilakukan pembatasan jumlah kombinasi

dengan mengelompokkan variabel-variabel independen menjadi 6 kategori yang akan menghasilkan 64 kombinasi. Pengelompokkan variabel independen dapat dilihat pada Tabel 3 Selanjutnya pada subbab berikutnya, penulisan kategori features akan ditulis dengan singkatannya saja, contoh *spotify features* akan disebut dengan SF.

**Tabel 3. Daftar Kategori Features**

Kategori Features	Variabel Independen yang Termasuk
Spotify Features (SF)	Spotify Rank, Spotify Velocity, Artist Popularity, Peak, Weeks on Chart, Song Age
Song Attribute (SA)	Danceability, Energy, Tempo, Valence, Duration
Shazam Rank (SR)	Shazam Rank
Airplay Rank (AR)	Airplay Rank
TikTok Rank (TR)	TikTok Rank
Other Chart Velocity (OV)	Shazam Velocity, Airplay Velocity, TikTok Velocity

Hasil *evaluation metric* dari pengujian ini dapat dilihat pada Tabel 4 Pada kolom kombinasi *features* (SF, SA, SR, AR, TR, dan OV), nilai 1 mengindikasikan bahwa variabel independen yang termasuk dalam kategori itu akan diikuti dalam pengujian model. Sebaliknya, variabel independen tidak akan diikuti apabila kolom bernilai 0. Karena baris data dari tabel ini sangatlah banyak, hasil tabel dibawah telah diurutkan berdasarkan rata-rata hasil A-R2 dari setiap *regressor* dari yang paling besar ke kecil serta diambil 8 peringkat teratas. Untuk memudahkan penyampaian analisa, ini juga hanya akan menggunakan *average* atau rata-rata dari *evaluation metric* tiap *regressor*. Diberikan juga index untuk pada setiap kombinasi memudahkan penyampaian kombinasi *features*.

**Tabel 4. Top 8 Hasil Pengujian Features Selection**

	SF	SA	SR	AR	TR	OV	Average	
							A-R2	MAE
21	1	0	1	0	1	0	90.559%	13.456
19	1	0	1	1	0	0	90.551%	13.487
23	1	0	1	0	0	0	90.550%	13.473
17	1	0	1	1	1	0	90.446%	13.509
22	1	0	1	0	0	1	90.324%	13.555
20	1	0	1	0	1	1	90.231%	13.561
18	1	0	1	1	0	1	90.201%	13.572
16	1	0	1	1	1	1	90.131%	13.560

### 5.3.1 Analisa Dampak Kategori Spotify Features

Terlihat pada Tabel 4 bahwa dari top 8 kombinasi kategori *features*, terdapat 8 kombinasi yang memiliki kolom SF atau *spotify features*

bernilai 1. Hal ini mengindikasikan bahwa penggunaan kategori SF memberikan pengaruh yang sangat signifikan terhadap model prediksi. Pada Tabel 5 diperlihatkan hasil *group by* kolom SF. Dari hasil *group by* kita dapat mengetahui rata-rata hasil *evaluation metric* dari semua kemungkinan yang memiliki nilai SF 1 dan semua kemungkinan yang memiliki nilai SF 0.

Dapat diketahui bahwa terdapat selisih nilai A-R2 sebesar 7.763% dan selisih nilai MAE sebesar -4.364. Dari informasi ini terlihat bahwa penggunaan variabel independen dari *spotify features* memberikan hasil positif yang signifikan pada model. Korelasi yang kuat ini disebabkan karena model yang dibuat memprediksi *ranking* sebuah lagu pada *chart* Spotify, sehingga tentu saja data dari *spotify features* memiliki korelasi yang kuat. Berdasarkan analisa yang sudah dipaparkan, kategori SF akan seterusnya digunakan dalam analisa dan pengujian model.

**Tabel 5. Hasil Group By Kategori Spotify Features Dari Pengujian Features Selection**

SF	SA	SR	AR	TR	OV	Average	
						A-R2	MAE
1	-	-	-	-	-	89.626%	13.764
0	-	-	-	-	-	81.863%	18.128

### 5.3.2 Analisa Dampak Kategori Song Attribute

Terlihat pada Tabel 4 bahwa dari top 8 kombinasi kategori *features*, terdapat 8 kombinasi yang memiliki kolom SA atau *song attribute* bernilai 0. Hal ini mengindikasikan bahwa penggunaan kategori SA memberikan pengaruh yang buruk terhadap model prediksi. Pada Tabel 6 diperlihatkan hasil *group by* kolom SA dengan nilai SF 1.

Dapat diketahui bahwa terdapat selisih nilai A-R2 sebesar -0.805% dan selisih nilai MAE sebesar 211. Dari informasi ini terlihat bahwa penggunaan variabel independen dari *song attribute* memberikan hasil negatif yang cukup berpengaruh pada model. Hal ini dapat disebabkan karena *trend* lagu *hit* zaman sekarang tidak terlalu terpengaruh dari kualitas sebuah lagu. Berdasarkan analisa yang sudah dipaparkan, kategori SA akan seterusnya tidak digunakan dalam analisa dan pengujian model.

**Tabel 6. Hasil Group By Kategori Song Attribute Dari Pengujian Features Selection**

SF	SA	SR	AR	TR	OV	Average	
						A-R2	MAE
1	1	-	-	-	-	89.223%	13.870
1	0	-	-	-	-	90.028%	13.659

### 5.3.3 Analisa Dampak Kategori Shazam Rank

Terlihat pada Tabel 4 bahwa dari top 8 kombinasi kategori *features*, terdapat 8 kombinasi yang memiliki kolom SR atau *shazam rank* bernilai 1. Hal ini mengindikasikan bahwa penggunaan kategori SR memberikan pengaruh yang baik terhadap model prediksi. Pada Tabel 7 diperlihatkan hasil *group by* kolom SR dengan nilai SF 1 dan nilai SA 0.

Dapat diketahui bahwa terdapat selisih nilai A-R2 sebesar 0.691% dan selisih nilai MAE sebesar -0.275. Dari informasi ini terlihat bahwa penggunaan variabel independen *shazam rank* memberikan hasil positif yang cukup berpengaruh pada model. Hal ini disebabkan karena adanya korelasi positif antara popularitas lagu di Shazam chart dan di *Spotify chart*, meskipun korelasinya tidak terlalu kuat. Berdasarkan analisa yang sudah dipaparkan, kategori SR akan seterusnya digunakan dalam analisa dan pengujian model.

**Tabel 7. Hasil Group By Kategori Shazam Rank Dari Pengujian Features Selection**

SF	SA	SR	AR	TR	OV	Average	
						A-R2	MAE
1	0	1	-	-	-	90.374%	13.522
1	0	0	-	-	-	89.683%	13.797

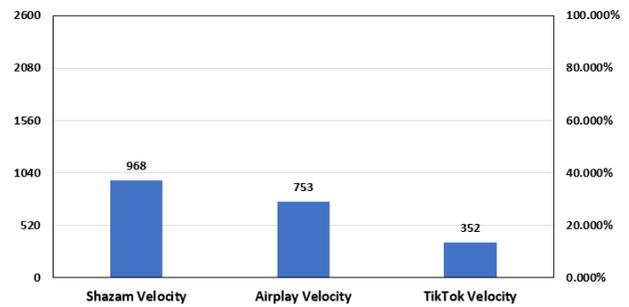
### 5.3.4 Analisa Dampak Kategori Other Chart Velocity

Terlihat pada Tabel 4 bahwa dari top 8 kombinasi kategori *features*, 4 kombinasi terbaik memiliki kolom OV atau *other chart velocity* bernilai 0. Hal ini mengindikasikan bahwa penggunaan kategori OV memberikan pengaruh yang buruk terhadap model prediksi. Pada Tabel 8 diperlihatkan hasil *group by* kolom OV dengan nilai SF 1, nilai SA 0, dan nilai SR 1.

**Tabel 8. Hasil Group By Kategori Other Chart Velocity Dari Pengujian Features Selection**

SF	SA	SR	AR	TR	OV	Average	
						A-R2	MAE
1	0	1	-	-	1	90.222%	13.562
1	0	1	-	-	0	90.527%	13.481

Dapat diketahui bahwa terdapat selisih nilai A-R2 sebesar -0.305% dan selisih nilai MAE sebesar 0.081. Dari informasi ini terlihat bahwa penggunaan variabel independen dari *other chart velocity* memberikan hasil negatif yang lemah pada model. Hal ini dapat disebabkan karena pergerakan rank antar minggu pada *chart* Shazam, Airplay, dan Tiktok memiliki pola yang kurang jelas. Alasan lain juga dapat disebabkan karena kurang banyaknya jumlah *data chart velocity* yang ter *join* dengan *chart* Spotify. Dapat dilihat pada Gambar 1 distribusi *data velocity* dari Shazam, Airplay, dan Tiktok secara berturut-turut adalah 968 (37.2% dari total data), 753 (28.9% dari total data), dan 353 (13.5% dari total data). Berdasarkan analisa yang sudah dipaparkan, kategori OV akan seterusnya tidak digunakan dalam analisa dan pengujian model.

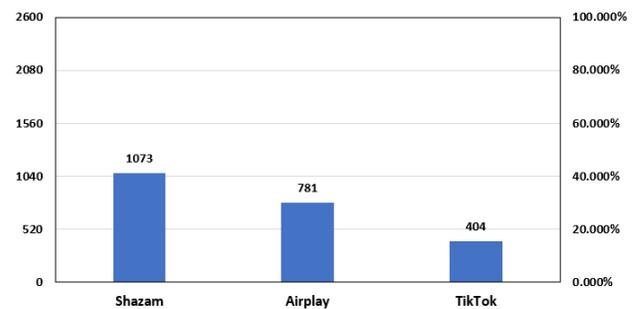


**Gambar 1. Distribusi Data Chart Yang Join Dengan Spotify**

### 5.3.5 Analisa Dampak Kategori Airplay Rank dan Tiktok Rank

Terlihat pada Tabel 4 bahwa dari top 8 kombinasi kategori *features*, 4 kombinasi terbaik tidak memiliki pola jelas yang mengindikasikan positif tidaknya penggunaan kategori airplay rank maupun tiktok rank. Pada 4 kombinasi teratas, terlihat juga perbedaan A-R2 dan MAE yang tidak banyak antar satu sama lain, yaitu 0.008-0.113% dan 0.017-0.053. Berdasarkan analisa yang sudah dipaparkan, pemakaian kedua kategori tersebut masih belum bisa ditentukan.

Meskipun begitu dari informasi yang didapatkan, dapat kita lihat bahwa korelasi dari kedua chart tersebut dengan *chart* Spotify dapat terbilang lemah. Alasan utama lemahnya korelasi ini dapat disebabkan dari sedikitnya jumlah data yang ter *join* dengan *chart* Spotify. Terlihat pada Gambar 2 bahwa distribusi data yang ter *join* untuk *chart* Airplay berjumlah 781 (30.0% dari total data), serta berjumlah 404 untuk *chart* Tiktok (15.5% dari total data). Hal ini disebabkan karena kurangnya persamaan dari jenis lagu yang berada pada *chart* Spotify dengan jenis lagu yang berada pada *chart* Airplay dan *chart* Tiktok.



**Gambar 2. Distribusi Data Chart Velocity**

Perbedaan jenis lagu pada *chart* Tiktok ini bisa disebabkan karena fungsi utama Tiktok yang tidak berfokus pada lagu. Berbeda dengan aplikasi Spotify dan Shazam, aksi utama *user* di Tiktok adalah pembuatan video. Sehingga *background sound*, yang terhitung pada *chart* Tiktok, hanyalah komponen sampingan dari pembuatan video. Ditambah lagi, *background sound* tidak selalu berupa lagu. Seringkali *user* hanya menggunakan instrumen musik, *original sound* yang berisi audio dari *user*, dan bahkan tidak menggunakan *sound* sama sekali. Hal ini membuat *chart* Tiktok tidak hanya berisikan lagu-lagu *official* saja, melainkan berisikan *sound* tidak *official* juga.

Sedangkan pada *chart* Airplay, perbedaan jenis lagu dapat disebabkan oleh perbandingan jumlah radio yang berbeda pada setiap *genre* lagu. Seperti yang disampaikan sebelumnya, Airplay adalah jumlah sebuah lagu diputarkan oleh radio pada suatu negara. Di

Amerika Serikat, radio-radio seringkali hanya berfokus pada sebuah *genre* saja, seperti *country radio*, *hip-hop/R&B*, dan *pop radio*. Pada 2021, jenis radio terbanyak adalah radio ber *genre country* dengan jumlah 2200 radio. Hal ini membuat mayoritas lagu pada *chart* Airplay merupakan *genre country* [6]. Berbeda dengan *chart* Airplay, konsumsi terbanyak musik *streaming* pada Amerika Serikat tahun 2021 dipegang oleh *genre hip-hop/R&B*. Lalu, diikuti oleh *genre rock, pop*, serta *genre country* pada peringkat ke-4 [7]. Hal ini membuktikan adanya perbedaan jenis lagu antara lagu yang diputar diradio dan lagu yang didengarkan pada *streaming platform*.

### 5.3.6 Pemilihan Kombinasi Features

Dari hasil analisa diatas, diketahui kategori mana yang berpengaruh baik dan yang harus kita gunakan dalam pengujian model kedepannya. Kategori yang akan digunakan adalah *spotify features* (SF) dan *shazam rank* (SR). Sedangkan kategori *song attribute* (SA) dan *other chart velocity* (OV) tidak akan digunakan karena terbukti memperburuk kinerja model. Masih terdapat ketidakpastian dalam pemakaian kategori *airplay rank* (AR) dan *tiktok rank* (TR). Oleh karena itu, semua kombinasi dengan kategori AR dan TR (nomor kombinasi 21, 19, 23, dan 17 pada Tabel 4) akan dicoba pada tahap *parameter tuning* disubbab selanjutnya. Pemakaian kategori AR dan TR akan ditentukan dari kombinasi model yang memberikan hasil paling optimal di tahap *parameter tuning*.

## 5.4 Pengujian Parameter Tuning

Pengujian ini akan dilakukan pada metode MLR, PR, GBT, dan RF menggunakan GridSearchCV untuk mencari *parameter* terbaik. Karena proses GridSearchCV dengan *parameter* yang banyak menghabiskan waktu yang lama, maka akan dilakukan pembatasan jumlah *parameter*. Daftar *parameter* yang digunakan dapat dilihat pada Tabel 9.

**Tabel 9. Daftar Parameter Yang Akan Digunakan**

Metode	Parameter
MLR	fit_intercept, normalize
PR	degree, include_bias
GBT	n_estimators, subsample, learning_rate, min_samples_leaf, max_depth
RF	n_estimators, max_features, max_depth

Nilai *cross validation* atau CV yang akan digunakan pada GridSearchCV adalah 5 dan 10. Sehingga akan dilakukan percobaan dengan menggunakan *parameter default*, *parameter tuning* dengan CV bernilai 5, dan *parameter tuning* dengan CV nilai 10. Dari hasil 3 percobaan CV ini, akan dipilih salah satu model dengan hasil terbaik. Setelah itu, percobaan ini akan dilakukan pada setiap *regressor* (MLR, PR, GBT, dan RF) dan setiap kombinasi *features* (nomor kombinasi 21, 19, 23, dan 17 pada Tabel 4). Pada Tabel 10 dapat dilihat hasil dari pengujian ini.

**Tabel 10. Hasil Pengujian Parameter Tuning**

SF+SR+AR+TR				
	MLR	PR	GBT	RF
A-R2	87.13%	91.08%	92.356%	93.133%
MAE	16.03	13.68	12.225	11.687
SF+SR+AR				
	MLR	PR	GBT	RF
A-R2	87.102%	91.886%	92.724%	92.916%
MAE	16.041	12.549	11.767	11.803
SF+SR+TR				
	MLR	PR	GBT	RF
A-R2	87.110%	91.040%	92.541%	93.094%
MAE	15.954	13.669	11.917	11.820
SF+SR				
	MLR	PR	GBT	RF
A-R2	87.130%	92.242%	92.236%	93.087%
MAE	16.031	12.131	12.478	11.694

## 5.5 Model Selection dan Analisa Secara Keseluruhan

Terlihat pada Tabel 10 bahwa model dengan hasil terbaik diperoleh oleh model dengan metode RF, CV bernilai 5, dan kombinasi *features* yang menggunakan SF, SR, AR, dan TR. *Parameter* yang digunakan adalah *max\_depth* bernilai 10, *max\_features* bernilai 4, dan *n\_estimators* bernilai 1000. Hasil *evaluation metric* yang didapatkan adalah nilai A-R2 sebesar 93.133% dan nilai MAE sebesar 11.687. Oleh karena itu, model ini lah yang terpilih sebagai best model serta akan digunakan dalam pembuatan aplikasi pada tahap selanjutnya.

Sebelumnya, pengujian model secara umum pada Tabel 1 mendapatkan model terbaik dengan metode GBT serta nilai A-R2 sebesar 90.861% dan MAE sebesar 13.191. Jika dibandingkan dengan best model yang telah diperoleh, terdapat kenaikan nilai A-R2 sebesar 2.272% dan penurunan nilai MAE sebesar 1.504. Hal ini membuktikan bahwa pemilihan *splitting ratio*, *features*, dan *parameter tuning* memberikan hasil yang positif terhadap kinerja model, mengingat adanya kenaikan pada nilai A-R2 dan penurunan pada nilai MAE. Selain itu, model terbaik yang sebelumnya dipegang oleh GBT juga berubah menjadi RF setelah dilakukan proses pemilihan *splitting ratio*, *features*, dan *parameter tuning*.

Pada Tabel 11 disajikan rata-rata hasil *evaluation metric* dari setiap *regressor* dari 2 model yang sudah di tuning. Model pertama hanya menggunakan *chart* Spotify saja. Sedangkan model kedua menggunakan *chart* Spotify ditambah dengan ketiga *chart* lainnya, yaitu Spotify, Shazam, Airplay, dan TikTok. Terlihat bahwa penggunaan ketiga *chart* lainnya memberikan dampak positif bagi model prediksi, dengan adanya peningkatan nilai A-R2 sebesar 0.951% dan penurunan nilai MAE sebesar 0.129. Akan tetapi, dampak yang diberikan tidak terlalu signifikan.

**Tabel 11. Perbandingan Hasil *Evaluation Metric* Antara Model Dengan *Spotify Chart* Dan Model Dengan Semua *Chart***

[	SR	AR	TR	Average	
				A-R2	MAE
1	0	0	0	89.854%	13.595
1	1	1	1	90.805%	13.466

*Best model* yang diperoleh memiliki kinerja yang sangat baik. Hal ini dilihat dari nilai *adjusted r-squared* sebesar 93.133% yang mengindikasikan bahwa model prediksi memiliki keakuratan yang sangat tinggi. Selain itu, nilai MAE juga memberi tahu bahwa rata-rata deviasi error dari model ini bernilai 11.687, dimana hal ini hanyalah 5.87% dari *maximum error* (200). Akan tetapi, jika dilihat pada penggunaannya dalam pengujian aplikasi, deviasi error dengan nilai 11.687 adalah nilai yang cukup besar. Jika sebuah data memiliki *actual value* atau *next week rank* bernilai 2, maka, berdasarkan MAE, hasil prediksi kurang lebih akan bernilai 13. Oleh karena itu, meskipun model ini sudah sangat baik, tetapi model ini masih belum bisa memberikan hasil prediksi yang sangat akurat kepada user.

## 6. KESIMPULAN

Secara keseluruhan, metode terbaik dalam melakukan prediksi peringkat mingguan lagu pada penelitian ini adalah *random forest*. Meskipun begitu, *Gradient boosting tree* dan *polynomial regression* masih termasuk *regressor* yang bekerja dengan sangat baik pada penelitian ini dan memiliki hasil *evaluation metric* yang bersaing dengan *regressor random forest*. Disisi lain, *Regressor* [ pada penelitian ini, meskipun masih tidak sebaik 3 *regressor* lainnya.

*Splitting ratio* terbaik dalam melakukan prediksi peringkat mingguan lagu pada penelitian ini adalah rasio dengan perbandingan 8:2. Untuk penggunaan *dataset*, kategori *song attribute* dan *other chart velocity*, terbukti memperburuk kinerja model. Sedangkan kategori *spotify features* dan *shazam rank* memberikan dampak yang sangat baik bagi model. Akan tetapi, kategori *tiktok rank* dan *airplay rank* tidak memberikan dampak positif maupun negative yang jelas. Meskipun begitu, terbukti bahwa kedua kategori ini memiliki relasi yang sangat lemah terhadap model.

Model terbaik yang didapatkan dari penelitian ini memiliki nilai *adjusted r-squared* sebesar 93.133% dan nilai *mean absolute error* sebesar 11.687. Model ini menggunakan metode *random forest*, *cross validation* bernilai 5, dan variabel independen dari *spotify chart*, *shazam chart*, *airplay chart*, dan *tiktok chart*. Parameter yang digunakan adalah *max\_depth* bernilai 10, *max\_features* bernilai 4, dan *n\_estimators* bernilai 1000.

## 7. SARAN

Untuk penelitian selanjutnya, dapat melakukan penelitian serupa dengan menggunakan *chart* musik lain, contohnya *chart Youtube*. Peneliti juga bisa melakukan pengumpulan data dengan skala yang lebih besar. Serta dapat melakukan *parameter tuning* dengan jumlah parameter dan jumlah kandidat yang lebih banyak.

## 8. REFERENSI

- [1] Araujo, C.V.S., Cristo, M.A.P., & Giusti, R. 2019. Predicting music popularity on streaming platforms. *Revista de Informatica Teorica e Aplicada - RITA*, 27(4), 108-117. DOI=10.5753/sbcm.2019.10436.
- [2] Brown, I. & Mues, C. 2012. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3), 3446-3453. DOI=10.1016/j.eswa.2011.09.033.
- [3] Chai, T. & Draxler, R. R. 2014. Root mean square error or mean absolute error. *The journal geoscientific model development*, 7, 1525-1534. DOI=10.5194/gmdd-7-1525-2014.
- [4] Chartmetric. Chartmetric api documentation. URI=https://api.chartmetric.com/apidoc/
- [5] Georgieva, E., Suta, M., & Burton, N. 2018. *Hitpredict: Predicting hit songs using spotify data*. Stanford University. URI=http://cs229.stanford.edu/proj2018/report/16.pdf.
- [6] Gotting, M. C. 2021, June 29. Leading radio formats in the United States as of June 2021, by number of stations. Statista. URI=https://www.statista.com/statistics/252230/top-radio-formats-in-the-us-by-number-of-stations/
- [7] Gotting, M. C. 2022, May 10. Distribution of streamed music consumption in the United States in 2021, by genre. Statista. URI=https://www.statista.com/statistics/475667/streamed-music-consumption-genre-usa/
- [8] International Federation of the Phonographic Industry. 2021. *IFPI issues global music report 2021*. URI=https://www.ifpi.org/ifpi-issues-annual-global-music-report-2021/.
- [9] Karch, J. 2019. *Improving on adjusted r-squared*. URI=http://dx.doi.org/10.31234/osf.io/v8dz5.
- [10] Karydis, I., Gkiokas, A., Katsouros, V., & Iliadis, L. 2018. Musical track popularity mining *Dataset: Extension & experimentation*. *Neurocomputing*, 280(1), 76-85. DOI=10.1016/j.neucom.2017.09.100.
- [11] Miles, J. 2014. R Squared, Adjusted R Squared. Wiley StatsRef: Statistics Reference Online. DOI=10.1002/9781118445112.stat06627.
- [12] Montgomery, D. C., Peck, E. A., & Vining, G. G. 1982. *Introduction to linear regression analysis*. DOI=10.1111/insr.12020\_10.
- [13] Rawlings, J. O., Dickey, D. A., & Pantula, S. G. 2006. *Applied regression analysis: A research tool* (2nd ed.). Germany: Springer New York.
- [14] Rencher, A. C. & Schaalje, G. B. 2008. *Linear models in statistics* (2nd ed.). Germany: Wiley.
- [15] Singhi, A. & Brown, D. G. 2014. *Hit song detection using lyric features alone*. University of Waterloo. URI=https://ismir2014.ismir.net/LBD/LBD12.pdf.
- [16] Spotify. *Spotify for developers documentation*. URI=https://developer.spotify.com/documentation/.
- [17] Spotify Technology S.A. 2021. *2020 annual report*. URI=https://investors.spotify.com/financials/default.aspx.
- [18] Suryanto, A. A. & Muqtadir, A. 2019. *Penerapan metode mean absolute error (MAE) dalam algoritma regresi linear untuk prediksi produksi padi*. *Saintekbu: Jurnal Sains dan Teknologi*, 11(1), 78-83. DOI=10.32764/saintekbu.v1i1.298