

Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Fitur Mode Audio Spotify

Lukhia Britanthia Christina Tanujaya^{a,1}, Bambang Susanto, ^{a,2}, Asido Saragih^{a,3}

^a *Departemen of Mathematics and Science Data-Universitas Kristen Satya Wacana, Gedung Y Jl. Diponegoro 52-60, Salatiga 50711, Indonesia*

¹ 662017003@student.uksw.edu; ² bambang.susanto@uksw.edu; ³ asido.saragih@uksw.edu

INFORMASI ARTIKEL	ABSTRAK
Diterima : 08 – 11 – 2020 Direvisi : 11 – 12 – 2020 Diterbitkan : 31 – 12 – 2020	Studi ini membandingkan kemampuan dari metode regresi logistik dan <i>random forest</i> dalam melakukan klasifikasi fitur mode. Fitur mode ini merupakan fitur yang terdapat di dalam data fitur audio. Secara keseluruhan, data ini berisikan data dari musik atau lagu yang dirilis di platform Spotify yang di dalamnya terdapat berbagai fitur dari masing-masing musik. Dalam melakukan studi ini, metode regresi logistik dan metode <i>random forest</i> ini diterapkan dalam bahasa pemrograman Python. Setelah dilakukannya studi ini dapat disimpulkan bahwa metode <i>random forest</i> dapat melakukan klasifikasi yang lebih baik walaupun dengan selisih yang cukup dekat. Karena kedua metode ini adalah metode yang baik dalam melakukan klasifikasi. Fitur penting yang ditampilkan oleh <i>random forest</i> juga memberikan hasil yang lebih memuaskan, karena fitur yang dihasilkan memang fitur yang berkaitan dengan fitur mode dan sesuai dengan teori musik.
Kata Kunci: fitur-fitur penting fitur mode klasifikasi <i>random forest</i> regresi logistik	
	

I. Pendahuluan

Menurut Jamalus, yang dituliskan dalam buku Seni Musik Jilid 1 untuk SMK, musik sendiri adalah suatu hasil karya seni bunyi dalam bentuk lagu atau komposisi musik, yang mengungkapkan pikiran dan perasaan penciptanya melalui unsur-unsur musik, yaitu irama, melodi, harmoni, bentuk/struktur lagu dan ekspresi sebagai satu kesatuan.[1] Mode adalah salah satu unsur terpenting yang terdapat dalam struktur hubungan antar *tone* yang merupakan rangkaian dasar musik. Mode yang banyak digunakan dalam musik ialah mayor dan minor. Perbedaan unsur penyusun mode mayor dan mode minor menghasilkan efek suara yang berbeda.[2], [3]

Terdapat beberapa metode yang dapat digunakan untuk klasifikasi, diantaranya adalah metode regresi logistik dan metode *random forest*. Regresi logistik adalah metode yang paling sering digunakan untuk memodelkan data variabel respon yang bersifat biner. *random forest* juga seringkali digunakan dalam prediksi dan klasifikasi, karena tingkat akurasi dari prediksinya yang tinggi berdasarkan pada banyaknya pohon.

Dalam melakukan klasifikasi mode mayor dan mode minor, penulis menggunakan metode regresi logistik dan *random forest*. Dari kedua metode ini akan dibandingkan metode yang paling baik dalam melakukan klasifikasi yang dilihat dari nilai akurasi, presisi, *recall*, dan *f1-score*. Selain itu, penulis juga ingin melihat variabel yang paling penting dari klasifikasi ini. Tahap yang dilakukan pada penelitian ini yaitu memisahkan data menjadi data latih dan data uji. Kemudian menerapkan metode regresi logistik dan *random forest*, dan membandingkan hasil dari kedua metode tersebut.

II. Metode

Dalam melakukan klasifikasi terdapat langkah-langkah penting, yaitu *learning* menggunakan data latih dan *testing* menggunakan data uji.

A. Pengolahan Data

Data yang digunakan dalam penelitian ini adalah data fitur audio dari Spotify yang rilis dari rentang waktu antara 2009 – 2019. Dari data ini kemudian dipisahkan antara data latih dengan data uji dengan persentase data latihnya 70% dan data ujinya 30% dari data asli. Kemudian dalam menerapkan metode regresi logistik dan *random forest*, data yang digunakan perlu normalisasi atau di transformasi terlebih dahulu. Untuk kasus regresi logistik, transformasi yang digunakan yaitu normalisasi min-max dengan bentuk umumnya yaitu

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

(1)

Untuk *random forest* transformasi yang digunakan adalah transformasi *standard scaler* dengan bentuk umumnya yaitu

$$z = \frac{x-u}{s} \quad (2)$$

dengan x adalah data sampel, u adalah *mean* dari sampel data latih, dan s adalah simpangan baku dari sampel data uji.

B. Klasifikasi

Klasifikasi adalah teknik untuk mengelompokkan data berdasarkan karakteristik dari data tersebut. Klasifikasi dapat dilakukan manual maupun dengan bantuan *machine learning*. Pada *machine learning* sudah menyediakan banyak bahasa pemrograman untuk klasifikasi, diantaranya yaitu dengan regresi logistik, *random forest*, *naïve bayes*, *support vector machine*, *fuzzy*, dan masih banyak lagi. Untuk mengetahui klasifikasi terbaik dapat dilihat dari akurasi, presisi, *recall*, dan *f1-score* yang didapatkan dari *confusion matrix* pada Tabel 1.[4]

Tabel 1. Confusion Matrix

		Respon (pengamatan; <i>observed</i>)	
		Ada : $Y = 1$	Tidak ada : $Y = 0$
Prediksi Respons	Ada $\hat{Y} = 1$	Positif Benar n_{TP} a	Positif Palsu n_{FP} B
	Tidak ada $\hat{Y} = 0$	Negatif Palsu n_{FN} c	Negatif Benar n_{TN} d

di mana,

n_{TP} : banyaknya subjek yang terdeteksi positif benar, artinya baik hasil maupun yang sebenarnya sama-sama positif.

n_{FP} : banyaknya subjek yang terdeteksi positif palsu, artinya hasilnya menunjukkan positif sedangkan seharusnya negatif.

n_{FN} : banyaknya subjek yang terdeteksi negatif palsu, artinya hasilnya menunjukkan negatif sedangkan seharusnya positif.

n_{TN} : banyaknya subjek yang terdeteksi negatif benar, artinya baik hasil maupun yang sebenarnya sama-sama menunjukkan negatif.

Setelah didapatkan hasil dari *confusion matrix*, dapat dilakukan juga perhitungan untuk *confusion matrix* seperti berikut[5] :

Akurasi, merupakan perbandingan dari subjek yang diidentifikasi benar dengan jumlah semua subjek.

$$accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (3)$$

Precision, merupakan perbandingan subjek yang memiliki hasil positif benar dengan jumlah hasil positif benar dan positif palsu.

$$precision = \frac{TP}{TP+FP} \quad (4)$$

Recall, merupakan perbandingan subjek yang memiliki hasil positif benar dengan jumlah hasil positif benar dan negatif palsu

$$recall = \frac{TP}{TP+FN} \quad (5)$$

F-score merupakan ukuran akurasi dari sebuah tes. *F1-score* sendiri merupakan perbandingan rata-rata presisi dan *recall* yang kemudian dituliskan seperti rumus dibawah ini.

$$f1\ score = \frac{2*precision*recall}{precision+recall} \quad (6)$$

Namun, setelah penelitian ini dilakukan, penulis menemukan bahwa rumus-rumus diatas hasilnya berbeda dengan perhitungan menggunakan *Python*. Perbedan ini terdapat pada rumus untuk presisi dan *recall*. Dari penelitian ini akan terlihat bahwa penulisan rumusnya tertukar, dimana rumus untuk presisi menjadi rumus untuk *recall*. Karena itu kedua rumus ini diperbaiki menjadi,

Recall, merupakan perbandingan subjek yang memiliki hasil positif benar dengan jumlah hasil positif benar dan positif palsu.

$$recall = \frac{TP}{TP+FP} \quad (7)$$

Precision, merupakan perbandingan subjek yang memiliki hasil positif benar dengan jumlah hasil positif benar dan negatif palsu

$$Precision = \frac{TP}{TP+FN} \quad (8)$$

C. Regresi Logistik

Metode regresi seringkali digunakan untuk menganalisis data untuk menunjukkan hubungan antara variabel respon dengan variabel penjelasnya. Dan akhir-akhir ini, regresi logistik lebih digemari. Perbedaan regresi logistik dengan regresi linear terdapat pada variabel responnya yang memiliki sifat biner atau dikotomik.[6]

Model dari regresi logistik berganda[4] adalah seperti berikut :

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon \quad (9)$$

di mana,

Y = variabel respon

X = variabel penjelas

β = konstanta atau intersep

β_i = koefisien X , untuk $i=1,2,\dots,k$

ε = residu atau galat

Dalam regresi logistik, untuk menguji hipotesa dan membuat interval kepercayaan parameter ε diasumsikan[7] :

- Nilai rata-rata dari galat adalah nol
- Variansi dari galat adalah konstan (homoskedastik)
- Tidak terjadi autokorelasi pada galat
- Galat berdistribusi normal

Oleh karena itu, dalam model regresi logistik terkadang parameter galat tidak dituliskan. Apabila nilai galatnya tidak nol pengaruhnya tidak signifikan dengan model karena nilai parameter sebenarnya tidak dapat diketahui.

D. Random Forest

Menurut Polamuri (2017), *Random forest* merupakan suatu algoritma yang sering digunakan pada klasifikasi data dengan jumlah besar karena tingkat akurasi dari prediksinya yang tinggi, dan berdasarkan pada banyaknya pohon.[8] Pembentukan pohonnya dilakukan secara acak yang kemudian akan digabungkan. Menurut Breiman (2001) dan Breiman & Cutler (2003), prosedur untuk melakukan *random forest* adalah

- Lakukan pengambilan sampel acak berukuran n dengan pengembalian. Tahap ini adalah tahapan *bootstrap*.
- Dengan menggunakan sampel *bootstrap*, pohon dibangun sampai mencapai ukuran maksimum (tanpa pemangkasan). Pembangunan pohon dilakukan dengan menerapkan *random feature selection* pada setiap proses pemilihan, yaitu k variabel penjelas dipilih secara acak.
- Ulangi langkah 1 dan 2, sehingga terbentuk sebuah hutan yang terdiri atas beberapa pohon.[9]

Karena menggunakan ansambel *decision tree*, *random forest* tidak dapat menentukan signifikansi setiap variabel dan hanya dapat menunjukkan tingkat kepentingan variabel. Sandri & Zuccolotto (2008) menuliskan rumus yang digunakan untuk menentukan tingkat kepentingan variabel. Misalkan terdapat q variabel penjelas dengan $h = 1,2,\dots,q$, maka *Mean Decrease in Gini* (MDG) mengukur tingkat kepentingan variabel penjelas X_h sebagai berikut[10] :

$$MDG_h = \frac{1}{k} \sum_t [d(h, t)I(h, t)] \quad (10)$$

dengan

k = banyaknya pohon dalam *random forest*

$d(h, t)$ = besar penurunan indeks *Gini* untuk variabel penjelas X_h pada simpul t

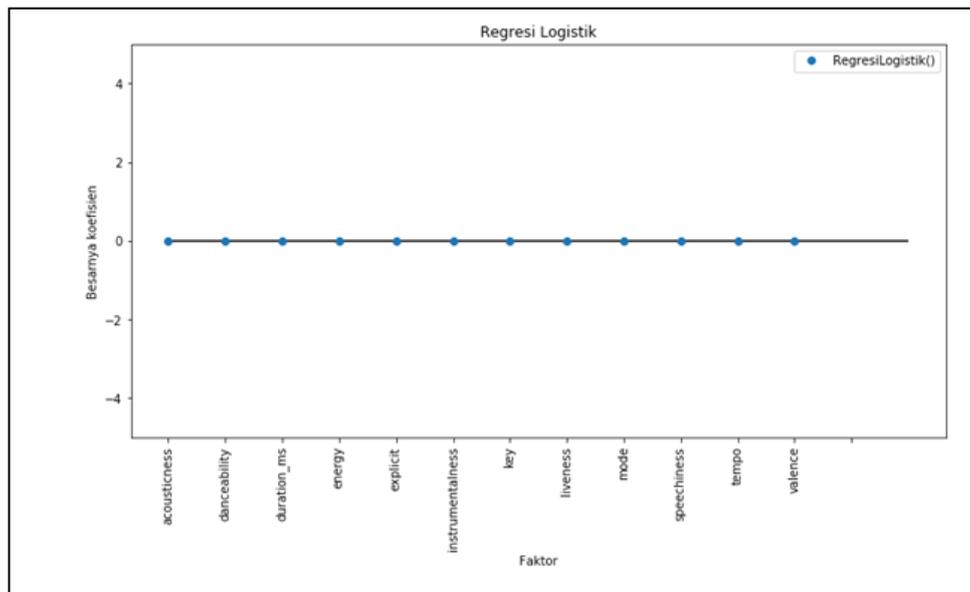
$$I(h, t) = \begin{cases} 1; & X_h \text{ memilah simpul } t \\ 0; & \text{selainnya} \end{cases}$$

III. Hasil dan Pembahasan

Sebelumnya, data audio ini diunggah terlebih dahulu ke dalam Python. Data ini terdiri dari 21789 lagu. Kemudian variabel-variabel dari data ini dibatasi, yakni dengan menghilangkan variabel *artists, id, name, popularity, release date, dan year*. Variabel respon dari data ini adalah variabel dari fitur mode yang terdiri dari mode mayor (1) dan mode minor (0). Setelah itu data dibagi menjadi data latih dan data uji dengan persentase data latihnya adalah 70% dari data asli, dan persentase data ujinya adalah 30% dari data asli. Tahap selanjutnya yaitu mengolah data ini untuk regresi logistik maupun *random forest*.

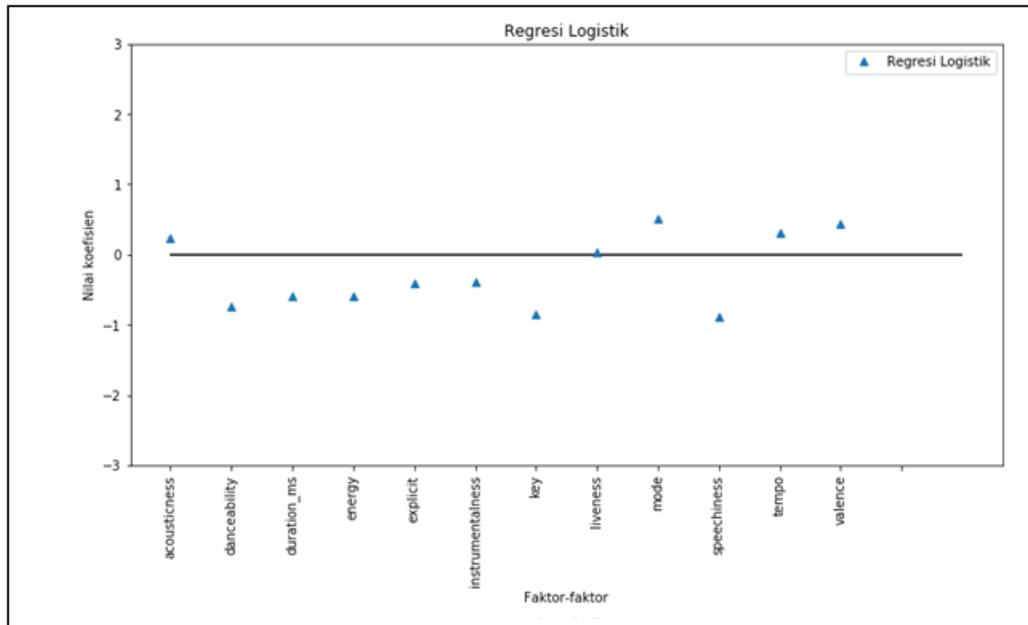
A. Regresi Logistik

Setelah dilakukan regresi logistik yang pertama, hasil yang ditampilkan belum berhasil. Nilai akurasi untuk data latih dari model regresi logistik sebesar 0.6593, sedangkan nilai akurasi untuk data ujinya sebesar 0.6404. Kemudian hasil dari model regresi logistik ini ditampilkan dalam bentuk grafik plot untuk melihat variabel apa saja yang menjadi variabel penjelas.



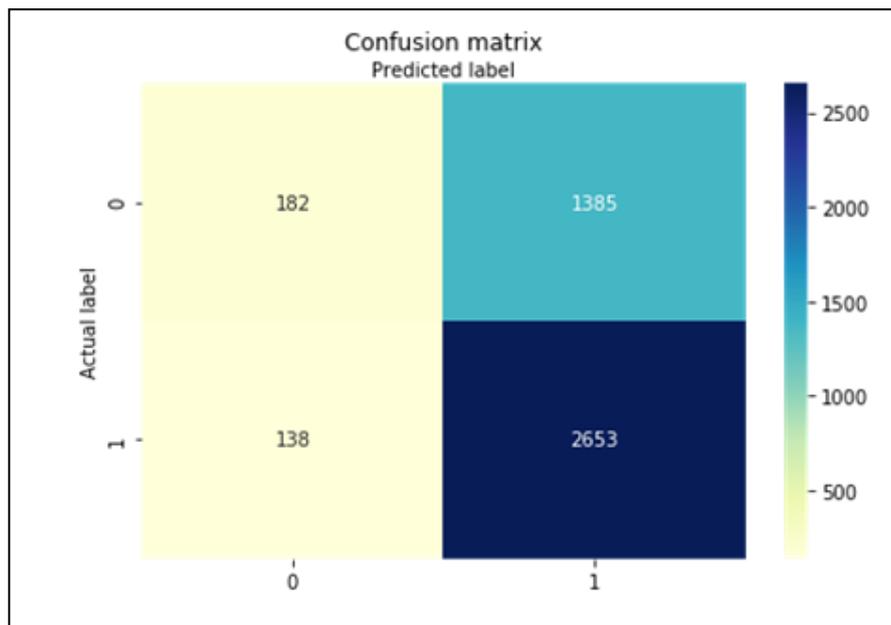
Gambar 1. Model Regresi Logistik 1

Ketidakberhasilan dari model regresi ini dapat dilihat dari Gambar 2, di mana besarnya koefisien yang ditunjukkan untuk masing-masing variabel bernilai 0. Untuk memperbaiki *error* ini, data yang akan dimodelkan ini perlu ditransformasi terlebih dahulu. Transformasi data yang digunakan untuk regresi logistik ini seperti yang sudah dituliskan di atas yaitu menggunakan normalisasi *min-max* dengan bentuk umumnya seperti pada Persamaan (1). Setelah ditransformasi, grafik plot yang dihasilkan akan menjadi seperti berikut.



Gambar 2. Model Regresi Logistik dengan Transformasi

Berdasarkan Gambar 3, persebaran koefisien dari variabel memang sudah menyebar. Namun, hanya terdapat dua variabel yang mencapai angka 1, yaitu variabel dari fitur *speechiness* dan fitur *key*. Variabel lainnya masih berada di sekitar atau mendekati angka 0. Selanjutnya, dilakukan klasifikasi dengan bantuan *confusion matrix* untuk menunjukkan dari hasil klasifikasi menggunakan regresi logistik.



Gambar 3. Confusion Matrix Regresi Logistik

Berdasarkan Gambar 4, dapat diketahui bahwa jumlah lagu yang terdeteksi sebagai mode minor dan benar sebanyak 182 lagu. Lagu yang terdeteksi sebagai mode minor namun sebenarnya adalah mayor sebanyak 1385 lagu, dan ini adalah jumlah yang cukup banyak. Untuk yang terdeteksi sebagai mayor namun ternyata adalah minor ada sebanyak 138 lagu. Dan yang terdeteksi sebagai lagu bermode mayor dan benar ada sebanyak 2653. Dari hasil *confusion matrix* ini terlihat bahwa dengan metode regresi logistik cukup kesulitan untuk mendeteksi lagu bermode minor. Kemudian dilanjutkan dengan menghitung akurasi, presisi, *recall*, dan *f1-score* dari klasifikasi menggunakan regresi logistik ini.

Tabel 2. Perhitungan Confusion Matrix Regresi Logistik

	precision	recall	f1-score	support
0	0.57	0.12	0.19	1567
1	0.66	0.95	0.78	2791
accuracy			0.65	4358
macro avg	0.61	0.53	0.48	4358
weighted avg	0.63	0.65	0.57	4358

akurasi: 0.6505277650298302

Hasil perhitungan dari Tabel 2 dapat dibuktikan dengan cara manual, yaitu:

$$accuracy = \frac{TP+TN}{TP+FP+FN+TN} = \frac{182+2653}{182+1385+138+2653} = 0.650528$$

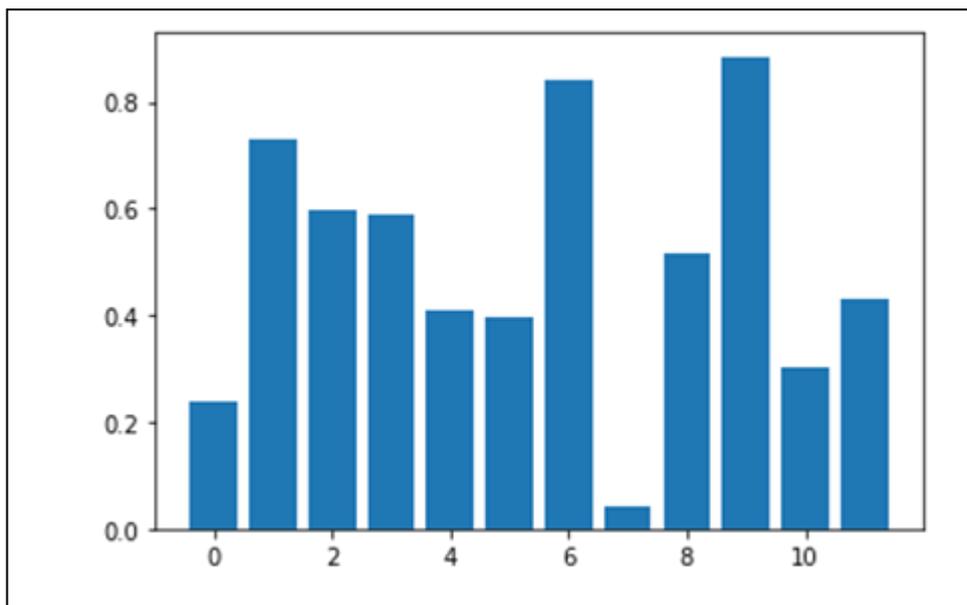
$$precision = \frac{TP}{TP+FN} = \frac{182}{182+138} = 0.5687 = 0.57$$

$$recall = \frac{TP}{TP+FP} = \frac{182}{182+1385} = 0.116146 = 0.12$$

$$f1\ score = \frac{2*precision*recall}{precision+recall} = \frac{2*0.116146*0.5687}{0.116146+0.5687} = 0.192888 = 0.19$$

$$error = \frac{FP+FN}{TP+FP+FN+TN} = \frac{1385+138}{182+1385+138+2653} = 0.349472 = 0.35$$

Dari hasil perhitungan manual ini terdapat perbedaan di hasil perhitungan untuk *recall* dan *precision*. Dari hasil, dapat di lihat bahwa hasil perhitungan manual menggunakan rumus dari sumber literasi yang didapat oleh penulis tertukar, di mana rumus untuk *recall* menjadi rumus perhitungan untuk *precision*. Oleh karena itu, dalam skripsi ini akan ditukar dan hasil perhitungan manual di atas menggunakan rumus yang sudah diperbaiki. Dapat di lihat bahwa hasil akurasi dari klasifikasi mode menggunakan metode regresi logistik ini sebesar 0.650528. Untuk nilai akurasi disini tidak terlalu baik. Pada umumnya, klasifikasi ini bisa dianggap baik apabila nilai akurasinya mendekati 90%. Kemudian untuk fitur-fitur yang membantu memodelkan mode ini dalam regresi logistik didapatkan hasil seperti pada gambar di bawah ini.

**Gambar 4. Variable Importance Regresi Logistik**

Keterangan dari Gambar 5 :

```
Feature: 0, Score: 0.239195
Feature: 1, Score: 0.730488
Feature: 2, Score: 0.597931
Feature: 3, Score: 0.587480
Feature: 4, Score: 0.410260
Feature: 5, Score: 0.396710
Feature: 6, Score: 0.840491
Feature: 7, Score: 0.041827
Feature: 8, Score: 0.516964
Feature: 9, Score: 0.884778
Feature: 10, Score: 0.303694
Feature: 11, Score: 0.431424
```

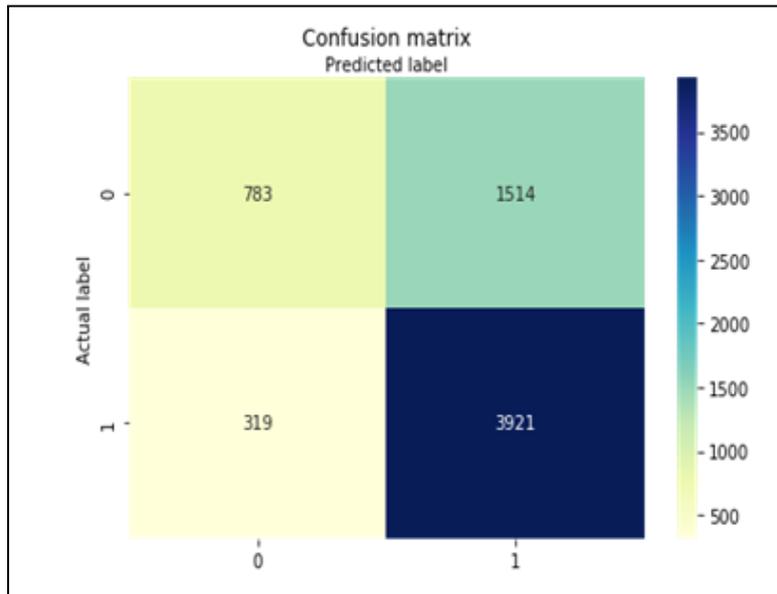
Hasil dari Gambar 5 apabila dikaitkan dengan fakta tentang unsur musik dan deskripsi dari arti fitur-fitur, maka didapatkan beberapa gagasan. **Pertama**, *speechiness* sebagai fitur terpenting ini menurut artinya adalah rekaman audio yang sebagian besar berisi kata-kata, seperti puisi, cerita, percakapan, atau dalam lagu misalnya *rap*. Memang beberapa *track* bisa saja mengandung musik, biasanya hanya sebagai iringan. Rekaman audio yang sebagian besar berisi kata-kata biasanya memilih jenis musik pengiring yang disesuaikan dengan cerita atau kata-kata yang disampaikan. Tetapi, apabila dibandingkan dengan fakta tentang unsur-unsur pembentuk musik, maka fitur ini tidak cocok. Ini dikarenakan fitur ini bukanlah yang mempengaruhi mode dari musik tetapi fitur ini membutuhkan musik dengan mode tertentu, sehingga relasi yang ada bukan berasal dari fitur yang mempengaruhi tetapi fitur yang membutuhkan.

Kedua, fitur *key* adalah fitur terpenting kedua yang berhubungan dengan fitur mode. Seperti yang sudah diketahui bahwa kunci sebuah lagu ada yang merupakan kunci mayor maupun minor. Oleh karena itu, masuk akal bila fitur ini penting bagi fitur mode karena merupakan salah satu unsur yang ikut menentukan mode dari suatu musik. **Ketiga**, hubungan antara fitur *danceability* dengan fitur mode yang ditunjukkan oleh hasil cukup tinggi. Fitur ini adalah fitur yang menuliskan seberapa cocok sebuah musik untuk ditarikan atau mengiringi tarian. Tarian memiliki banyak sekali nuansa, baik yang positif maupun negatif. Karena itu, tarian ini membutuhkan musik pengiring yang sesuai dengan nuansanya. Selain itu, musik juga bisa memberikan warna sendiri yang mempengaruhi tarian.

Selanjutnya, untuk fitur *duration* cukup sulit menemukan hubungan antara fitur ini dengan fitur mode. Untuk fitur *energy* sendiri memang terlihat adanya keterkaitan di mana musik yang memiliki tingkat *energy* yang tinggi misalnya cepat biasanya adalah lagu atau musik yang bernuansa positif. dan lagu atau musik yang bernuansa positif ini seperti yang kita tahu biasanya memiliki mode mayor. Kemudian untuk fitur yang secara jelas tertulis menjelaskan tentang suasana dari lagu, yaitu fitur *valence* justru berada di urutan ke-7. Begitu pula dengan fitur *tempo* yang biasanya berkontribusi memperkuat suasana dari lagu justru berada di urutan ke-10. Oleh karena itu, model regresi logistik belum cukup bisa membuktikan bahwa fitur-fitur ini merupakan fitur terpenting bagi fitur mode.

B. Random Forest

Setelah dilakukan pemisahan data menjadi data latih dan data uji, akan dilakukan klasifikasi menggunakan *random forest*. Dalam *random forest*, data langsung ditransformasi menggunakan menggunakan *standard scaler*. Banyaknya pohon estimasi yang digunakan, yaitu 1000. Selanjutnya hasil klasifikasi dari *random forest* ditunjukkan melalui *confusion matrix* di bawah ini.



Gambar 5. Confusion Matrix Random Forest

Berdasarkan Gambar 6, dapat diketahui bahwa jumlah lagu yang terdeteksi sebagai mode minor dan benar sebanyak 783 lagu. Lagu yang terdeteksi sebagai mode minor namun sebenarnya adalah mayor sebanyak 1514 lagu. Untuk yang terdeteksi sebagai mayor namun ternyata adalah minor ada sebanyak 319 lagu. Dan yang terdeteksi sebagai lagu bermode mayor dan benar ada sebanyak 3921. Dari hasil *confusion matrix* ini terlihat bahwa dengan metode regresi logistik cukup kesulitan untuk mendeteksi lagu bermode minor. Selanjutnya dilakukan juga perhitungan berdasarkan *confusion matrix* di atas.

Tabel 3. Perhitungan Confusion Matrix Random Forest

	precision	recall	f1-score	support
0	0.71	0.34	0.46	2297
1	0.72	0.92	0.81	4240
accuracy			0.72	6537
macro avg	0.72	0.63	0.64	6537
weighted avg	0.72	0.72	0.69	6537

akurasi: 0.7195961450206517

Hasil perhitungan dari Tabel 3 dapat dibuktikan dengan cara manual, yaitu :

$$accuracy = \frac{TP+TN}{TP+FP+FN+TN} = \frac{783+3921}{783+1514+319+3921} = 0.719596 = 0.72$$

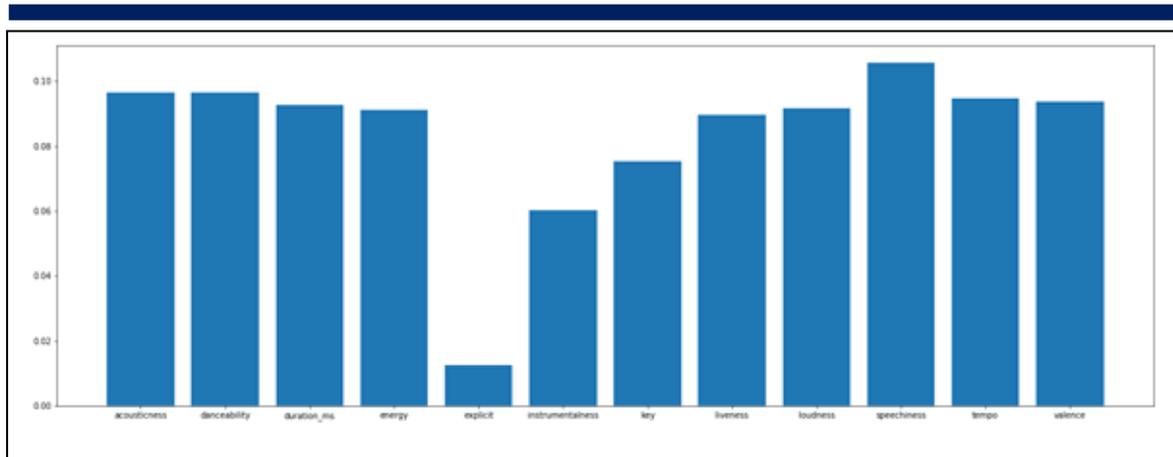
$$precision = \frac{TP}{TP+FN} = \frac{783}{783+319} = 0.7105263$$

$$recall = \frac{TP}{TP+FP} = \frac{783}{783+1514} = 0.340879 = 0.34$$

$$f1\ score = \frac{2*precision*recall}{precision+recall} = \frac{2*0.340879*0.7105263}{0.340879+0.7105263} = 0.460723 = 0.46$$

$$error = \frac{FP+FN}{TP+FP+FN+TN} = \frac{1514+319}{783+1514+319+3921} = 0.280404 = 0.28$$

Nilai akurasi dari data latih sebesar 0.9986805117319718 dan akurasi dari data uji sebesar 0.7085819183111519. Kemudian diberikan fitur-fitur penting yang membantu dalam model *random forest*.



Gambar 6. Variable Importance Random Forest

Penjelasan lengkap dari Gambar 7 seperti berikut :

var	feature	importance
9	speechiness	0.105736
0	acousticness	0.096377
1	danceability	0.096359
10	tempo	0.094757
11	valence	0.093761

Dari Gambar 7 yang disajikan terlihat bahwa fitur *speechiness* kembali menempati posisi pertama. Ini berarti *speechiness* kembali menjadi fitur yang paling membantu memodelkan dalam random forest. Setelah *speechiness* disusul oleh *acousticness*, *danceability*, *tempo*, dan *valence*. Hasil yang diberikan dapat diterima karena sebagian besar fitur dalam kenyataannya jelas memiliki hubungan dengan fitur mode. *Tempo*, pada kenyataannya apabila sebuah lagu bertempo lambat biasanya bernuansa negatif yang berarti sama dengan mode minor. *Valence* adalah fitur yang mewakili nuansa lagu di mana untuk lagu yang valensinya mendekati angka 0 adalah lagu yang bernuansa negatif, sama halnya dengan mode minor. Sebuah lagu bila cocok untuk ditarikan berarti lagu itu bisa membawa suasana sekitar menjadi lebih semarak atau positif, dan ini sama dengan mode mayor yang biasanya memiliki nuansa positif.

C. Hasil Perbandingan Regresi Logistik dan Random Forest

Pertama, akan dibandingkan hasil dari *confusion matrix*, namun yang sudah dalam bentuk perhitungan.

Tabel 4. Perbandingan Perhitungan Confusion Matrix

	Regresi Logistik (minor)	Random Forest (minor)
precision	0.57	0.71
recall	0.12	0.34
f1-score	0.19	0.46
accuracy	0.6505277650298302	0.7195961450206517

Berdasarkan Tabel 4, hasil klasifikasi dari *random forest* lebih tinggi dibandingkan hasil dari regresi logistik, baik dari segi presisi sampai ke nilai akurasinya. Selanjutnya, dibandingkan juga hasil dari grafik fitur penting yang dapat membantu klasifikasi di kedua metode yang akan disajikan dalam bentuk tabel.

Tabel 5. Perbandingan Variable Importance

Regresi Logistik		Random Forest	
Fitur	Nilai	Fitur	Nilai
<i>speechiness</i>	0.884778	<i>speechiness</i>	0.105736
<i>key</i>	0.840491	<i>acousticness</i>	0.096377
<i>danceability</i>	0.730488	<i>danceability</i>	0.096359
<i>duration</i>	0.597931	<i>tempo</i>	0.094757
<i>energy</i>	0.587480	<i>valence</i>	0.093761

Dari Tabel 5, dapat diketahui bahwa *random forest* juga memberikan hasil yang lebih baik. Ini dikarenakan tiap fitur yang berada di peringkat 5 ke atas merupakan fitur dari audio yang memiliki hubungan dengan fitur mode dalam teori musik. Sedangkan di regresi logistik terdapat fitur *duration*, yang mana apabila dihubungkan dengan arti dari durasi menurut data fitur ini maka hasilnya tidak bisa diterima. Seperti yang kita ketahui, lagu bernuansa negatif tidak memiliki ciri yang berhubungan dengan durasi, baik durasi yang lama maupun yang pendek. Hasil dari regresi logistik maupun *random forest* menunjukkan bahwa pada peringkat 1 dan peringkat 3 dari kedua metode ini diisi oleh fitur yang sama, yaitu *speechiness* dan *danceability*.

IV. Kesimpulan

Berdasarkan uraian dari hasil dan pembahasan penelitian ini, dapat disimpulkan bahwa:

1. Klasifikasi menggunakan metode *random forest* lebih baik daripada menggunakan metode regresi logistik. Hal ini ditunjukkan dari nilai akurasi *random forest* yang lebih tinggi dibandingkan nilai akurasi dari regresi logistik. *Random forest* mampu melakukan 71% benar dalam mengklasifikasi mode musik atau lagu. Semakin banyak *tree* yang digunakan, maka semakin baik pula akurasinya. Selisih nilai akurasi *random forest* dengan regresi logistik memang cukup dekat karena keduanya adalah metode klasifikasi yang baik.
2. Dengan bantuan regresi logistik dan *random forest* sebagai *classifier*, dapat ditunjukkan fitur-fitur penting yang membantu model untuk klasifikasi. Baik regresi logistik maupun *random forest* menunjukkan bahwa fitur *speechiness* dan *danceability* merupakan fitur terpenting. Semakin sedikit alunan musik yang terkandung dalam sebuah lagu atau audio, semakin tidak mungkin untuk digunakan sebagai pengiring tarian. Hal itu menyebabkan tidak ada mode yang bisa terdeteksi dalam audio tersebut. Karena itu, kedua fitur ini tetap ada di dua metode yang berbeda. Namun, *random forest* memberikan hasil yang lebih baik, karena semua fiturnya secara teoritis memang berhubungan dengan adanya mode minor maupun mayor.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada pihak-pihak yang telah membantu penerbitan paper ini, yaitu kepada yang pertama dan terutama Tuhan Yesus Kristus, karena berkat karunia dan bimbingan-Nya segala proses yang sudah penulis lalui hingga tahap ini dapat terlewati. Selanjutnya kepada Bapak Dr. Adi Setiawan, M.Sc. selaku Dekan Fakultas Sains dan Matematika, Bapak Dr. Bambang Susanto, M.S. dan Bapak Asido Saragih, S.Pd, M.Sc. selaku Pembimbing untuk segala bimbingan, kesabaran dan ide-ide yang berharga untuk penulis, Bapak, Ibu dosen, dan staf FSM Matematika yang telah memberikan ilmu dan pengetahuan yang berharga kepada penulis selama penulis studi di FSM-UKSW, dan keluarga terkasih yang selalu mendoakan dan memberikan semangat kepada penulis, serta para *reviewer* dan panitia dari penerbit yang memberikan kesempatan yang luar biasa bagi penulis.

Daftar Pustaka

- [1] M. Muttaqin, *Seni Musik Klasik Jilid 1 untuk SMK*, Jilid 1., vol. 1. Jakarta: Direktorat Pembinaan Sekolah Menengah Kejuruan, Direktorat Jenderal Manajemen Pendidikan Dasar dan Menengah, Departemen Pendidikan Nasional, 2008.
- [2] L. Fang, J. Shang, and N. Chen, "Perception of western musical modes: A Chinese study," *Front. Psychol.*, vol. 8, no. OCT, pp. 1–8, 2017, doi: 10.3389/fpsyg.2017.01905.
- [3] I. Press, "The Affective Character of the Major and Minor Modes in Music Author (s): Kate Hevner Source : The American Journal of Psychology , Jan ., 1935 , Vol . 47 , No . 1 (Jan ., 1935), Published by : University of Illinois Press Stable URL : <http://www.jsto>," vol. 47, no. 1, pp. 103–118, 1935.
- [4] A. . Fallis, *Analisis Regresi Logistik*, vol. 53, no. 9. 2013.
- [5] H. D. Anggana, "Penerapan Model Klasifikasi Regresi Logistik, Support Vector Machine , Classification and Regression Tree Terhadap Data Kejadian Difteri Di Provinsi Jawa Barat," *Euclid*, vol. 5, no. 2, p. 20, 2018, doi: 10.33603/e.v5i2.1121.
- [6] D. W. Hosmer and S. Lemeshow, "Applied Logistic Regression." pp. 1–375, 2000, doi: 10.1074/jbc.272.33.20373.
- [7] H. Wibowo, W. Rahayu, P. S. Matematika, U. N. Jakarta, and R. L. Biner, "OVERDISPERSI PADA

-
- REGRESI LOGISTIK BINER MENGGUNAKAN METODE BETA BINOMIAL,” p. 53, 2007.
- [8] R. A. Haristu, “PENERAPAN METODE RANDOM FOREST UNTUK PREDIKSI WIN RATIO PEMAIN PLAYER UNKNOWN BATTLEGROUND,” Universitas Sanata Dharma, 2019.
- [9] L. Breiman, “Random Forest,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1017/CBO9781107415324.004.
- [10] M. Sandri and P. Zuccolotto, “A bias correction algorithm for the gini variable importance measure in classification trees,” *J. Comput. Graph. Stat.*, vol. 17, no. 3, pp. 611–628, 2008, doi: 10.1198/106186008X344522.
- [11] Breiman L, Cutler A. 2003. Manual on Setting Up, Using, and Understanding Random Forest V4.0. [terhubung berkala]. http://oz.berkeley.edu/users/breiman/Using_random_forests_v4.0.pdf.
- [12] Hilbe, J. M. (2011). *Logistic Regression Lotion Logistic Regression Regression*. 2, 113.
- [13] Jamalul. (1988.). *Buku pengajaran musik melalui pengalaman musik panduan pengajar/ Jamalul*. Jakarta :: Proyek Pengembangan Lembaga Pendidikan Tenaga Kependidikan,.
- [14] Nirwana. S.R.A. 2015. Regresi Logistik Multinomial dan Penerapannya dalam Menentukan Faktor yang Berpengaruh pada Pemilihan Program Studi di Jurusan Matematika UNM. Skripsi. Universitas Negeri Makassar. Makassar.
- [15] Polamuri, Saimadhu. How Random Forest Algorithm Works In Machine Learning <https://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learning/>
- [16] Williams G. (2011) Random Forests. In: Data Mining with Rattle and R. Use R. Springer, New York, NY