# PREDIKSI KEPOPULERAN LAGU BERDASARKAN TANGGA LAGU BILLBOARD MENGGUNAKAN DECISION TREE DAN K-MEANS

**Desta Gumilar**[1*], **Tacbir Hendro Pudjiantoro**[1], **Rezki Yuniarti**[1]
[1] Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Jenderal Achmad Yani
Jl. Terusan Jendral Sudirman, Cimahi, Jawa Barat, 40285
[*]Email: gumilar92@gmail.com

**Abstrak**

*Billboard telah menjadi sumber terpercaya untuk peringkat popularitas lagu selama 60 tahun terakhir, dan sebagian besar label rekaman mengacu pada peringkat yang diberikan Billboard. Lagu hits biasanya tidak hanya dipengaruhi oleh lirik dan artis yang menyanyikannya, beberapa lagu hits dipengaruhi juga oleh faktor sepeti artis, genre, label rekaman dan lain sebagainya. Namun jika label rekaman dapat memprediksi sendiri apakah sebuah lagu dapat masuk ke dalam peringkat yang ada di Billboard tentunya akan sangat membantu. Dalam tangga lagu terdapat atribut yang dipertimbangkan seperti artis, judul, genre dan lainnya sehingga kombinasi berbagai atribut tersebut menjadi suatu pola dalam mengelompokkan sebuah lagu dalam tangga lagu. Penelitian ini bertujuan membangun sistem yang dapat memprediksi apakah suatu lagu dapat dikategorikan menjadi hits atau tidak menggunakan Decision Tree dan K-Means.*

***Kata kunci**: decision tree, k-means, lagu, prediksi*

## 1. PENDAHULUAN

Industri lagu belakangan ini mengalami perubahan yang sangat signifikan. Para penikmat lagu kini cenderung mengunjungi atau membeli *file* lagu secara *online* dibanding pergi ke sebuah toko. Lagu yang disukai atau sedang *hits* cukup mudah didapat, karena banyak penelitian yang dilakukan seperti rekomendasi musik maupun sistem pengambilan keputusan yang dapat membantu penikmat lagu dalam menemukan lagu yang sesuai. Penikmat lagu bisa mengakses situs seperti Billboard yang menyediakan tanggal lagu apabila mengalami kesulitan dalam mengetahui lagu yang sedang *hits*.

Billboard telah menjadi sumber terpercaya untuk peringkat popularitas lagu selama 60 tahun terakhir, dan sebagian besar label rekaman mengacu pada peringkat yang diberikan Billboard (Cibils, Meza & Ramel 2015). Lagu hits biasanya tidak hanya dipengaruhi oleh lirik dan artis yang menyanyikannya, beberapa lagu hits dipengaruhi juga oleh faktor sepeti artis, genre, label rekaman dan lain sebagainya, seperti Slipknot yang masih menempati tangga lagu meski hanya digemari oleh kalangan tertentu. Penelitian sebelumnya mengemukakan sebuah lagu dikatakan *hits* karena memiliki beberapa karakteristik, karakteristik tersebut di antaranya memiliki PV, *beat* lagu memiliki level antara sedang dan cepat, panjang lagu antara 3,36 sampai 4,12 menit, lagu baru seharusnya tidak *featuring* dan dirilis pada bulan September, Oktober dan November, banyaknya penghargaan yang dimiliki seorang artis juga mempengaruhi kemungkinan lagu masuk jajaran *hits* (Banpotsakun & Chongwatpol 2015). Penelitian tersebut bisa menjadi acuan bagi label apabila artis yang menyanyikan sebuah lagu sudah di ketahui. Apabila belum diketahui, penelitian tersebut kurang bisa menjadi acuan, sehingga sulit bagi label rekaman untuk mengetahui lagu yang mereka keluarkan masuk pada peringkat Billboard. Oleh karena itu dibutuhkan suatu cara yang dapat memudahkan label rekaman dalam memprediksi sendiri lagu yang akan keluarkan.

Penelitian ini bertujuan membangun sistem yang dapat memprediksi apakah suatu lagu dapat dikategorikan menjadi *hits* atau tidak berdasarkan faktor seperti artis, album, *genre*, *featuring*, *soundtrack* dan label. Pembagian kelompok menjadi tiga kelas yaitu *top chart*, *middle chart*, dan *bottom chart* diharapkan dapat menambah keakuratan dalam memprediksi lagu karena pada penelitian sebelumnya mengemukakan bahwa dengan sepuluh pembagian kelompok, hasil keakuratan yang dicapai yakni sebesar 67% (Koenigstein, Shavitt & Zilberman 2009).

## 2. METODOLOGI

Metode penelitian yang dilakukan untuk sistem prediksi lagu ini dibagi menjadi beberapa tahap. Proses awal adalah pengumpulan data yang berhubungan dengan tangga lagu. Data yang di gunakan adalah data lagu dan riwayat tangga lagu pada situs Billboard yang nantinya di olah sehingga menghasilkan enam atribut yaitu artis, album, *genre*, *featuring*, *soundtrack*, label, posisi dan populer. Atribut artis terbagi menjadi *male solo*, *female solo*, *male group*, *female group* dan *band*. Album dibagi menjadi *single* atau *album*. Genre dibagi menjadi *blues*, *country*, *folk*, *jazz*, *pop*, *reggae*, *rock*, *R&B* dan sebagainya. *Featuring* dibagi menjadi ya atau tidak. Label dibagi menjadi *sony*, *EMI*, *colombia*, *atlantic* dan sebagainya.

### 2.1. Proses Data

Data yang digunakan sebagai masukan adalah data lagu dan riwayat tangga lagu yang berasal dari situs *Billboard* pada periode 2012-2015. Data tersebut dijadikan data latih untuk membentuk pohon keputusan yang akan membentuk syarat yang memiliki atribut judul lagu, artis, dan posisi. Sedangkan untuk data uji, dipilih data lagu pada periode lain dengan durasi satu tahun terakhir.

#### 2.1.1. Proses Data Masukan

Data lagu yang digunakan sebagai masukan kemudian dilakukan pemrosesan menggunakan Decision Tree dan K-means untuk mengetahui apakah lagu dapat dikategorikan menjadi *hits* atau tidak. Hasil prediksi tersebut digunakan untuk gambaran bagi label musik tentang lagu yang kemungkinan akan menjadi *hits*.

##### 2.1.1.1 Praproses

Pada tahap ini dilakukan pra-proses dengan langkah-langkah sebagai berikut:

1. Data Selection

   Pertama dilakukan pemilihan data yang akan digunakan untuk pemrosesan, data yang dipilih adalah data tangga lagu dari tahun 2012 sampai 2015 pada situs Billboard.
2. Cleaning

   Proses *cleaning*/pembersihan antara lain menghilangkan duplikasi data, memeriksa inkonsistensi data, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi) serta menambahkan data lain yang relevan
3. Transformation

   Tahap ini dilakukan pengubahan data menjadi format yang bisa digunakan untuk keperluan proses selanjutnya. Contoh data tangga lagu yang sudah diubah untuk proses prediksi dapat dilihat pada Tabel 1.

***Tabel 1. Contoh Data Latih Tangga Lagu***

| No | Artis | Album | Genre | Featuring | Soundtrack | Label | Posisi | Populer |
|---|---|---|---|---|---|---|---|---|
| 1 | Male Solo | Album | Soul | NO | YES | Columbia | 1 | Yes |
| 2 | Female Solo | Album | Trap | YES | NO | Capitol | 2 | Yes |
| 3 | Male Solo | Album | R&B | NO | NO | Columbia | 3 | Yes |
| 4 | Female Solo | Album | Electrohop | YES | NO | Virgin | 4 | Yes |
| 5 | Male Group | Album | Pop Rock | NO | NO | Interscope | 5 | Yes |
| 6 | Male Solo | Album | R&B | YES | NO | Atlantic | 6 | Yes |
| 7 | Band | Album | Pop | NO | NO | Sony | 7 | Yes |
| 8 | Female Solo | Album | R&B | NO | NO | Epic Record | 8 | Yes |
| 9 | Female Solo | Album | Dancepop | YES | NO | Republic | 9 | Yes |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 28 | Male Solo | Album | Synthpop | YES | NO | Interscope | 28 | No |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 29 | Band | Album | Pop Rock | NO | NO | Interscope | 29 | No |
| 30 | Male Solo | Album | R&B | YES | NO | RCA | 30 | No |

#### 2.1.1.2 Proses Prediksi

Data yang telah melalui tahap pra-proses selanjutnya dilakukan penghitungan dengan konsep e*ntropy* dan *information gain*. Konsep ini digunakan untuk menentukan *node* induk dan *node* daun dalam Decision tree (Pudjiantoro, Renaldi & Teogunadi 2011).

Dimulai dari *node* induk, harus dihitung terlebih dahulu *entropy* untuk semua data dengan menggunakan Persamaan (1).

$$Entropy(S) = -P_+ \log_2 P_+ - P_- log_2 P_- \quad (1)$$

Dimana :
S = ruang (data) sample yang digunakan untuk training.
P+ = jumlah yang bersolusi positif (mendukung) pada data sample untuk kriteria tertentu.
P- = jumlah yang bersolusi negatif (tidak mendukung) pada data sample untuk kriteria tertentu.

Tiap-tiap fitur mulai dari artis, album dan seterusnya terkecuali fitur posisi dihitung untuk mencari nilai *entropy* masing-masing. Hasil yang didapatkan dari semua perhitungan dapat dilihat pada Tabel 2.

***Tabel 2. Entropy Keseluruhan***

| Fitur | Nilai | Jumlah data | Yes | No | Entropy |
|---|---|---|---|---|---|
| | Entropy Awal | 30 | 10 | 20 | 0,91829 |
| Artis | Male solo | 10 | 4 | 6 | 0,97095 |
| | Female solo | 11 | 4 | 7 | 0,94566 |
| | Male group | 5 | 1 | 4 | 0,72193 |
| | Female group | 0 | 0 | 0 | 0 |
| Album | Band | 4 | 1 | 3 | 0,81128 |
| | Album | 25 | 10 | 15 | 0,97095 |
| | Single | 5 | 1 | 4 | 0,72193 |
| Genre | Soul | 4 | 3 | 1 | 0,81128 |
| | Pop | 3 | 1 | 2 | 0,9183 |
| | Trap | 3 | 1 | 2 | 0,9183 |
| | R&B | 4 | 3 | 1 | 0,81128 |
| | Electro hop | 2 | 1 | 1 | 1 |
| | Hip-hop | 3 | 1 | 2 | 0,9183 |
| | Pop rock | 3 | 1 | 2 | 0,9183 |
| | Dance pop | 3 | 1 | 2 | 0,9183 |
| | Synthpop | 2 | 0 | 2 | 0 |
| | Indie pop | 1 | 0 | 1 | 0 |
| | Folk rock | 1 | 0 | 1 | 0 |
| | Folk pop | 1 | 0 | 1 | 0 |
| | Electro pop | 1 | 0 | 1 | 0 |
| Featuring | Yes | 11 | 4 | 7 | 0,94566 |
| | No | 19 | 6 | 13 | 0,89974 |
| Soundtrack | Yes | 1 | 1 | 0 | 0 |
| | No | 29 | 9 | 20 | 0,89357 |
| Label | Columbia | 5 | 2 | 3 | 0,97095 |
| | Capitol | 2 | 2 | 0 | 0 |
| | Virgin | 2 | 1 | 1 | 1 |
| | Interscope | 5 | 1 | 4 | 0,72193 |
| | Atlantic | 1 | 1 | 0 | 0 |
| | Sony | 2 | 1 | 1 | 1 |
| | Epic | 2 | 1 | 1 | 1 |
| | Republic | 4 | 1 | 3 | 1 |
| | RCA | 2 | 0 | 2 | 0 |

| | | | | |
|---|---|---|---|---|
| Big Machine | 1 | 0 | 1 | 0 |
| EMI | 1 | 0 | 1 | 0 |
| Walt Disney | 1 | 0 | 1 | 0 |
| Inertia | 1 | 0 | 1 | 0 |
| Def Jam | 1 | 0 | 1 | 0 |

Setelah dilakukan pencarian *entropy* kemudian dihitung *gain* untuk setiap fitur dengan menggunakan Persamaan (2).

$$Gain(S, A) = Entropy(S) - \sum \frac{|S_v|}{S} Entropy(S_v) \quad (2)$$

Dimana :
S = ruang (data) sampel yang digunakan untuk *training*.
A = atribut.
V = suatu nilai yang mungkin untuk atribut A.
|Sv| = jumlah sample untuk nilai V.
|S| = jumlah seluruh sample data.
Entropy(Sv) = entropy untuk sample-sample yang memiliki nilai V

Berikutnya setelah pencarian *entropy* dan *gain* dilakukan proses pengelompokan lagu menggunakan metode K-Means, data latih yang digunakan yakni berdasarkan Tabel 3.1 yang sudah terlebih dahulu di normalisasi. Normalisasi dilakukan dengan cara mengubah data yang ada sehingga untuk selanjutnya bisa di proses. Atribut artis seperti *male solo*, *female solo*, *male group*, *female gorup*, dan *band* di normalisasi menjadi 1 sampai 5. *Album*, *Featuring* dan *Sountrack* yang memiliki dua nilai di normalisasi menjadi menjadi 1 dan 0. Sedangkan *Genre* dan *Label* yang memiliki banyak nilai, terlebih dahulu di generalisasi berdasarkan induknya seperti genre *soul* dan *neo soul* yang menginduk kepada genre R&B atau label *capitol* dan *republic* yang menginduk kepada label universal. Normalisasi dari atribut dapat dilihat pada Tabel 3.

***Tabel 3. Normalisasi Data Atribut***

| | | | | | |
|---|---|---|---|---|---|
| **Artis** | Male Solo | 1 | **Soundtrack** | YES | 1 |
| | Female Solo | 2 | | NO | 0 |
| | Male Group | 3 | **Genre** | Alternative | 1 |
| | Female Group | 4 | | Country | 2 |
| | Band | 5 | | EDM/Electronic | 3 |
| **Album** | Album | 1 | | Hip hop/rap | 4 |
| | Single | 0 | | Pop | 5 |
| **Label** | Sony | 1 | | R&B/Soul | 6 |
| | Universal | 2 | | Rock | 7 |
| | Warner | 3 | | Reggae | 8 |
| **Featuring** | YES | 1 | | Indie | 9 |
| | NO | 0 | | Lain lain | 0 |

Berdasarkan tabel normalisasi yang terbentuk, berikut 24 contoh data tangga lagu yang sudah diubah, seperti pada Tabel 4.

***Tabel 4. Normalisasi Data Tangga Lagu***

| No | Artis | Album | Genre | Featuring | Soundtrack | Label |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 6 | 0 | 1 | 1 |
| 2 | 2 | 1 | 4 | 1 | 0 | 2 |
| 3 | 1 | 1 | 6 | 0 | 0 | 1 |
| 4 | 2 | 1 | 3 | 1 | 0 | 2 |
| 5 | 3 | 1 | 5 | 0 | 0 | 2 |
| 6 | 1 | 1 | 6 | 1 | 0 | 3 |
| 7 | 5 | 1 | 5 | 0 | 0 | 1 |
| 8 | 2 | 1 | 6 | 0 | 0 | 1 |
| 9 | 2 | 1 | 5 | 1 | 0 | 2 |
| 10 | 1 | 1 | 6 | 0 | 0 | 2 |
| 11 | 1 | 1 | 5 | 1 | 0 | 1 |
| 12 | 5 | 1 | 5 | 0 | 0 | 2 |
| 13 | 2 | 1 | 5 | 0 | 0 | 2 |
| 14 | 3 | 1 | 5 | 0 | 0 | 3 |
| 15 | 3 | 0 | 4 | 0 | 0 | 1 |
| 16 | 1 | 1 | 4 | 1 | 0 | 2 |
| 17 | 3 | 1 | 9 | 1 | 0 | 1 |
| 18 | 2 | 0 | 3 | 0 | 0 | 2 |
| 19 | 1 | 1 | 0 | 0 | 0 | 1 |
| 20 | 2 | 0 | 3 | 0 | 0 | 2 |
| 21 | 2 | 0 | 5 | 0 | 0 | 3 |
| 22 | 1 | 1 | 6 | 0 | 0 | 1 |
| 23 | 5 | 1 | 5 | 0 | 0 | 2 |
| 24 | 3 | 1 | 0 | 0 | 0 | 1 |

Berdasarkan tabel yang terbentuk selanjutnya dipilih tiga pusat kelompok yaitu data ke-8, 16 dan data ke-24 kemudian dihitung jarak dari setiap data dengan mencari nilai terdekat dari titik pusat kelompok ke-1 (C1) hingga kelompok ke-3 (C3) yang telah ditentukan sebelumnya menggunakan Persamaan (3).

$$d_{ij} = \sqrt{\sum_{k=1}^{p}\{x_{ik} - x_{jk}\}^2} \quad (3)$$

Dimana :
$d_{ij}$ = Jarak objek antara objek i dan j
P = Dimensi data
$X_{ik}$ = Koordinat dari obyek i pada dimensi k
$X_{jk}$ = Koordinat dari obyek j pada dimensi k

Hasil keseluruhan dari perhitungan K-Means dapat dilihat pada Tabel 5.

***Tabel 5. Hasil Perhitung K-means***

| No | C1 | C2 | C3 | Terkecil | Kelompok |
|---|---|---|---|---|---|
| 1 | 1,4142 | 2,6458 | 6,4031 | 1,4142 | C1 |
| 2 | 2,4495 | 1 | 4,3589 | 1 | C2 |
| 3 | 1 | 2,4495 | 6,3246 | 1 | C1 |
| 4 | 3,3166 | 1,4142 | 3,4641 | 1,4142 | C2 |
| 5 | 1,7321 | 2,4495 | 5,099 | 1,7321 | C1 |
| 6 | 2,4495 | 2,2361 | 6,7082 | 2,2361 | C2 |
| 7 | 3,1623 | 4,3589 | 5,3852 | 3,1623 | C1 |
| 8 | 0 | 2,6458 | 6,0828 | 0 | C1 |

| | | | | | |
|---|---|---|---|---|---|
| 9 | 1,7321 | 1,4142 | 5,2915 | 1,4142 | C2 |
| 10 | 1,4142 | 2,2361 | 6,4031 | 1,4142 | C1 |
| 11 | 1,7321 | 1,4142 | 5,4772 | 1,4142 | C2 |
| 12 | 3,3166 | 4,2426 | 5,4772 | 3,3166 | C1 |
| 13 | 1,4142 | 1,7321 | 5,1962 | 1,4142 | C1 |
| 14 | 2,4495 | 2,6458 | 5,3852 | 2,4495 | C1 |
| 15 | 2,4495 | 2,6458 | 4,1231 | 2,4495 | C1 |
| 16 | 2,6458 | 0 | 4,6904 | 0 | C2 |
| 17 | 3,3166 | 5,4772 | 9,0554 | 3,3166 | C1 |
| 18 | 3,3166 | 2 | 3,4641 | 2 | C2 |
| 19 | 6,0828 | 4,2426 | 2 | 2 | C3 |
| 20 | 3,3166 | 2 | 3,4641 | 2 | C2 |
| 21 | 2,4495 | 2,2361 | 5,5678 | 2,2361 | C2 |
| 22 | 1 | 2,4495 | 6,3246 | 1 | C1 |
| 23 | 3,3166 | 4,2426 | 5,4772 | 3,3166 | C1 |
| 24 | 6,0828 | 4,6904 | 0 | 0 | C3 |

## 3. HASIL DAN PEMBAHASAN

Pengujian sistem dilakukan dengan menggunakan data latih sebanyak 250 data set dan data uji sebanyak 78 data set. Pengujian ini mencari kesesuaian kelas populer dari data latih terhadap data uji dan mencari kesesuaian kategori antara *top chart*, *middle chart*, atau *bottom chart* yang dicari berdasarkan pembentukan aturan yang telah terbentuk sebelumnya. Berdasarkan pengujian yang telah dilakukan, terdapat 73 data set dengan keluaran sesuai dengan masukan. Sedangkan 5 data set tidak sesuai dengan masukan.

## 4. KESIMPULAN

Penelitian ini menghasilkan sebuah sistem yang dapat memprediksi apakah suatu lagu dapat dikategorikan menjadi *hits* atau tidak menggunakan Decision Tree dan K-Means. Berdasarkan hasil pengujian pada 78 data set, Sistem dapat menghasilkan *presicion* sebesar 80%, *recall* sebesar 70% dan akurasi sebesar 93%.

Dalam penelitian ini, masih terdapat banyak kekurangan yang mungkin dapat dikembangkan untuk penelitian selanjutnya dengan topik yang serupa. Adapun saran dari hasil penelitian ini untuk penelitian selanjutnya yaitu:

a. Jumlah data lagu untuk data latih yang diolah diharapkan dapat lebih banyak.
b. Penambahan atribut atau faktor-faktor yang berpengaruh terhadap kepopuleran lagu untuk variasi data latih

## DAFTAR PUSTAKA

Banpotsakun, P & Chongwatpol, J 2015, 'Determining the Key Success Factors for hit songs in the Billboard Music Charts', *SAS Institute Inc*, no. 3368, pp. 1-12.

Cibils, C, Meza, Z & Ramel, G 2015, 'Predicting a Song's Path through the Billboard Hot 100', *CS*.

Koenigstein, N, Shavitt, Y & Zilberman, N 2009, 'Predicting Billboard Success Using Data-Mining in P2P Networks', *IEEE International Symposium on Multimedia*, no. 978-0-7695-3890-7.

Pudjiantoro, TH, Renaldi, F & Teogunadi, A 2011, 'Penerapan Data Mining Untuk Menganalisa Kemungkinan Pengunduran Diri Calon Mahasiswa Baru', *KNS&I*, no. 009.