

## IMPLEMENTASI DATA MINING UNTUK PREDIKSI TANGGA LAGU APLIKASI LANGIT MUSIK

Hanifah Ulfah Rabbani<sup>1</sup>, Dian Anubhakti<sup>2</sup>

<sup>1</sup>Mahasiswa Jurusan Sistem Informasi, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta

<sup>2</sup>Dosen Jurusan Sistem Informasi, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta

### ABSTRACT

Indonesia is a one of the countries with a large population that can be used as a market for various products, including online music player technology products targeting a young audience. One application that is trending right now is Langit Musik. A streaming music platform that is widely used by millennials. With pattern recognition techniques to see trends from a current and future song, hypothesis testing and search algorithms are needed. This paper aims to predict the top 30 hits, how LangitMusik can change the millennials music trends as well as the trend of streaming music technology.

**Keyword:** prediction, streaming online, music, millennial.

### 1. PENDAHULUAN

Tingginya minat masyarakat Indonesia dengan bidang teknologi berdampak pada salah satu aplikasi pemutar musik *online*. Generasi saat ini lebih memilih mendengarkan musik secara online melalui perangkat digital. Survey yg dilakukan Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), menampilkan bahwa **13.3%** dari populasi pengguna internet di Indonesia atau sekitar 35 juta orang Indonesia mendengarkan musik secara online (APJII, 2018). Berdasar infografis, sekitar 87% pengguna LangitMusik adalah generasi milenial (data LangitMusik, 2020). Generasi ini menjadi pasar terbesar yang mengkonsumsi aplikasi streaming musik. Dari berbagai macam lagu diaplikasi LangitMusik, para penikmat musik akan mencari music terpopuler Untuk mengetahui tingkat kepopuleran suatu lagu, biasanya para penikmat akan menentukan atribut seperti artist, album, genre, duet bahkan soundtrack dari movie. Dari berbagai macam atribut tadi, para penikmat musik bisa menentukan apakah suatu lagu tersebut masuk dalam kategori top popular atau tidak.

Penelitian sebelumnya berpendapat bahwa sebuah lagu dikatakan popular karena memiliki karakteristik, seperti PV, beat lagu memiliki level antara cepat dan sedang, panjang lagu berdurasi 3:36 hingga 4:12 menit, lagu baru seharusnya tidak berduet dan dirilis dibulan September, Oktober, November, banyaknya penghargaan yang dimiliki oleh sang artist (Banpotsakun & Chongwatpol 2015). Tapi hal tersebut tidak bisa dijadikan suatu acuan bagi para penikmat lagu untuk menentukan kepopuleran suatu lagu.

Salah satu metode yang bisa membantu prediksi adalah Algoritma C4.5. Algoritma C4.5 adalah algoritma klasifikasi data dengan teknik decision tree yang terkenal dan disukai karena memiliki kelebihan-kelebihan, seperti: dapat mengolah data numerik (kontinyu) dan diskret, dapat menangani nilai atribut yang hilang, menghasilkan aturan-aturan yang mudah diinterpretasikan dan tercepat di antara algoritma-algoritma yang menggunakan memori utama di komputer (Quinlan, 1993; Han et al., 2001; Berry et al., 1997; Ruggieri, 2001). Sedangkan Decision adalah struktur sederhana yang dapat digunakan sebagai pengklasifikasi. Referensi penting dalam pengerjaan aslinya adalah Classification and Regression Tree oleh Breiman et al. Pada pohon keputusan, masing-masing node internal (non-leaf) merepresentasikan sebuah variabel atribut (atribut prediksi atau fitur) dan masing-masing cabang merepresentasikan satu keadaan dari variabel ini. Pohon keputusan dibuat menggunakan himpunan dari data yang digunakan sebagai data pembelajaran (training dataset).

Oleh sebab itu, penelitian ini akan menggunakan metode Algoritma C4.5 dalam memprediksi kepopuleran musik berdasar tangga lagu dari LangitMusik sehingga bisa membantu para penikmat musik untuk menentukan kepopuleran suatu lagu

### 2. METODOLOGI PENELITIAN

Algoritma C4.5 juga menangani kasus yang tidak memiliki nilai untuk sebuah atau lebih atribut. Akan tetapi, atribut kelas hanya bertipe nominal dan tidak boleh kosong. Secara teoritis, (Emha Taufiq Luthfi, 2009) menyebutkan beberapa tahapan untuk mengaplikasikan algoritma C4.5, yakni:

1. Pilih atribut sebagai akar

---

<sup>1</sup> Korenspondensi Penulis : Hanifah Ulfah Rabbani : 085330055325, hanifahrabbani19@gmail.com

Untuk memilih atribut sebagai akar, didasarkan pada nilai gain tertinggi dari atribut yang ada. Rumus untuk menghitung nilai gain bisa dilihat di persamaan 1 berikut ini :

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * Entropy(S_i)$$

Keterangan:

S : Himpunan kasus A

n : Jumlah partisi atribut A

|S<sub>i</sub>| : Jumlah kasus pada partisi ke-i

|S| : Jumlah kasus dalam S

Sementara untuk perhitungan entropy, menggunakan persamaan 2 berikut ini:

$$Entropy(S) = \sum_{i=1}^n -p_i * \log^2 p_i$$

Keterangan:

S : Himpunan kasus

A : Fitur

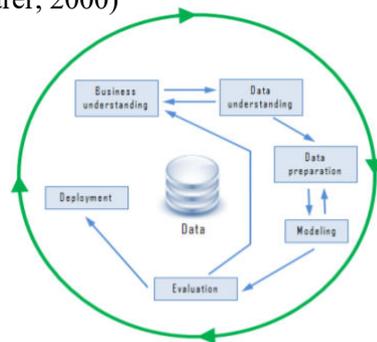
n : Jumlah partisi S

p<sub>i</sub> : Proporsi dari S<sub>i</sub> terhadap S

2. Buat cabang untuk tiap-tiap nilai
3. Bagi kasus dalam cabang
4. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Untuk Metode CRISP-DM terdiri dari 6 fase.

1. *Business Understanding* adalah pemahaman tentang substansi dari kegiatan data *mining* yang akan dilakukan, kebutuhan dari sisi bisnis seperti menentukan tujuan data *mining*.
2. *Data Understanding* adalah fase pengumpulan data awal, mempelajari data yang akan dipakai, mendeteksi subset yang menarik dari data untuk membuat hipotesa awal.
3. *Data Preparation* adalah fase dimana aktivitas yang dilakukan seperti memilih *table* dan *field* yang akan ditransformasikan kedalam *database* untuk bahan data *mining*
4. *Modelling* adalah fase untuk menentukan teknik data *mining* yang digunakan, tools data mining serta teknik data mining.
5. *Evaluation* adalah fase interpretasi terhadap hasil data mining yang ditunjukkan dalam proses pemodelan. Evaluasi dilakukan secara mendalam dengan tujuan menyesuaikan model yang didapat agar sesuai dengan tujuan dalam fase pertama.
6. *Deployment* adalah fase penyusunan laporan atau presentasi dari pengetahuan yang didapat dari evaluasi pada proses data mining (Shearer, 2000)



Gambar 1. Model CRISP-DM

Proses Data yang digunakan sebagai masukan adalah data play lagu dan riwayat tangga lagu dari LangitMusik pada periode 2020. Data tersebut dijadikan data latih untuk membentuk pohon. Data lagu yang digunakan sebagai masukan kemudian dilakukan pemrosesan menggunakan *Decision Tree* untuk mengetahui apakah lagu dapat dikategorikan top populer atau tidak. 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 tahun 2020. Data yang digunakan bersumber dari data LangitMusik dengan dengan jumlah kasus sebanyak 3,635,864 record, dan terdiri dari 10 atribut, seperti terlihat dalam Gambar 2.

SONG_NAME	ARTIST_NAME	ALBUM_NAME	GENRE_NAME	LABEL_NAME	GENDER
Mutiara Yang Hil...	Vina Panduwinata	Mutiara Yang Hilang (Single)	Indonesia - Pop	Virgo Multi Cipta	F
Teman Setiaku	BRIGHT TEA	"BEST IN HEART", Vol. 2	Indonesia - Pop	Believe International	M
Fly My Eagle	Anggun	"FLY MY EAGLE" - OST 'Pend...	Soundtrack - Local Movie	Miles Music	F
Pop That Lock (...)	Adam Lambert	"Trespassing" EP	Barat - Pop	Sony Music International	M
"Wetten, dass...?"	Pentatonix	"Wetten, dass...?" Medley	Barat - Pop	Sony Music International	H
Always Be My B...	Mariah Carey	#1 to Infinity	Barat - Pop	Sony Music International	F
My All	Mariah Carey	#1 to Infinity	Barat - Pop	Sony Music International	F
One Sweet Day	Mariah Carey feat. Boyz II Men	#1 to Infinity	Barat - Pop	Sony Music International	H
Hero	Mariah Carey	#1 to Infinity	Barat - Pop	Sony Music International	F
We Belong Tog...	Mariah Carey	#1 to Infinity	Barat - Pop	Sony Music International	F
Emotions	Mariah Carey	#1 to Infinity	Barat - Pop	Sony Music International	F
Someday (MTV ...)	Mariah Carey	#1 to Infinity	Barat - Pop	Sony Music International	F
Don't Forget Ab...	Mariah Carey	#1 to Infinity	Barat - Pop	Sony Music International	F
Love Takes Time	Mariah Carey	#1 to Infinity	Barat - Pop	Sony Music International	F
Vision of Love	Mariah Carey	#1 to Infinity	Barat - Pop	Sony Music International	F
I Dont Wanna C...	Mariah Carey	#1 to Infinity	Barat - Pop	Sony Music International	F
Honey	Mariah Carey	#1 to Infinity	Barat - Pop	Sony Music International	F
Dreamlover	Mariah Carey	#1 to Infinity	Barat - Pop	Sony Music International	F
Infinity	Mariah Carey	#1 to Infinity	Barat - Pop	Sony Music International	F
Touch My Body	Mariah Carey	#1 to Infinity	Barat - Pop	Sony Music International	F
Without You	Mariah Carey	#1 to Infinity	Barat - Pop	Sony Music International	F
Don't Forget Ab...	Mariah Carey	#1 to Infinity	Barat - Pop	Sony Music International	F
Emotion	Destinys Child	#1s	Barat - R&B	Sony Music International	F
Whenever You ...	Mariah Carey,Brian McKnight	#1s	Barat - Pop	Sony Music International	H

Gambar 2. Potongan dataset.

## 2. Data Cleaning

Proses *cleaning* adalah proses menghilangkan duplikasi data, memeriksa data yang tidak konsisten dan bersifat *noise* dari data yang terdapat di berbagai basis data yang mungkin berbeda format maupun platform yang kemudian diintegrasikan dalam satu database *datawarehouse*.

## 3. Data Transformation

Tahap ketiga adalah tahap *pre-processing*, dimana data akan ditransformasikan menjadi data yang siap untuk diolah menjadi data *mining*. Tujuannya untuk pengubahan data menjadi format yang bisa digunakan untuk keperluan proses selanjutnya.

## 4. Data Mining

Proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan metode Decision Tree Teknik. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

## 5. Data Evaluation

Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses Knowledge Discovery in Database (KDD) yang disebut *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

## 4. PEMBAHASAN

Untuk mendapatkan hasil penelitian maka dilakukan :

1. Menghitung jumlah kasus class YA dan class TIDAK serta nilai Entropy dari semua kasus. Kasus dibagi berdasarkan atribut pada Gambar 3 dengan jumlah kasus 3635864 record, class YA ada 2743486 record dan kelas TIDAK sebanyak 892378 record sehingga didapat entropy:

$$= (-2743486/3635864 \cdot \log_2(2743486/3635864)) + (-892378/3635864 \cdot \log_2(892378/3635864)) = \mathbf{0,803966601}$$

2. Menghitung nilai gain dari masing-masing atribut terlihat pada pada Gambar 3. dan pohon keputusan pada, sehingga didapat nilai gain:

$$= (0,80396660695015) - ((1720557/3635864) * 0,771726) - ((795793/3635864) * 0,832294) - ((225680/3635864) * 0,970769) = \mathbf{0,006456238}$$

KASUS		Jumlah	Yes	No
GENDER		3635864	2743486	892378
	Male - Solo	1720557	1330981	389576
	Female - Solo	795793	585934	209859
	Group	893834	691093	202741
	Duet	225680	135478	90202
GENRE		3635864	2588185	1047679
	POP	2292476	2243817	48659
	DANGDUT	550206	0	550206
	ROCK	151780	150455	1325
	Alternative	71748	71731	17
	RELIGI	159740	122182	37558
	Other	409914	0	409914
LABEL		1451224	1333260	117964
	SONY	486098	485516	582
	TRINITY	233197	233197	0
	MUSICA	228120	228120	0
	WARNER	169543	169543	0
	UNIVERSAL	334266	216884	117382

Gambar 3. Tabel Kasus berdasarkan atribut.

KASUS		Jumlah	Yes	No	Entropy	Gain	SPLIT INFO	GAIN RATIO
GENDER		3635864	2743486	892378	0,803967	0,006456	1,737065693	0,00371675
	Male - Solo	1720557	1330981	389576	0,771726			
	Female - Solo	795793	585934	209859	0,832294			
	Group	893834	691093	202741	0,77243			
	Duet	225680	135478	90202	0,970769			
GENRE		3635864	2588185	1047679	0,866321			
	POP	2292476	2243817	48659	0,148267	0,735188	1,799683794	0,408509513
	DANGDUT	550206	0	550206	0			
	ROCK	151780	150455	1325	0,072249			
	Alternative	71748	71731	17	0,003195			
	RELIGI	159740	122182	37558	0,78683			
LABEL		1451224	1333260	117964	0,406694			
	SONY	486098	485516	582	0,013347	0,186837	2,221769848	0,084093916
	TRINITY	233197	233197	0	0			
	MUSICA	228120	228120	0	0			
	WARNER	169543	169543	0	0			
	UNIVERSAL	334266	216884	117382	0,935103			

Gambar 4. Tabel Hasil perhitungan nilai Entropy dan Gain.

### Hasil Pengujian dengan WEKA

Pada tahapan penelitian ini dilakukan pengujian tingkat akurasi dengan metode data mining, metode yang digunakan adalah algoritma C4.5. Pengujian dilakukan dengan menggunakan *tools WEKA* dengan tujuan untuk lebih meningkatkan kepastian akan tingkat akurasi yang diperoleh. Tahapan ini menggunakan operator *cross validation* yang berfungsi untuk mengestimasi tingkat akurasi performa berdasarkan pembelajaran model dari berbagai algoritma yang terdapat pada *WEKA*. Terlihat pada Gambar 5. metode pengujian algoritma dengan C4.5 dan hasil Decision Tree pada aplikasi *WEKA* Gambar 6.

```

Time taken to build model: 69.35 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      922421          90.8505 %
Incorrectly Classified Instances    92897           9.1495 %
Kappa statistic                    0.8121
Mean absolute error                0.1369
Root mean squared error            0.2613
Relative absolute error            28.0484 %
Root relative squared error        52.8932 %
Total Number of Instances         1015318

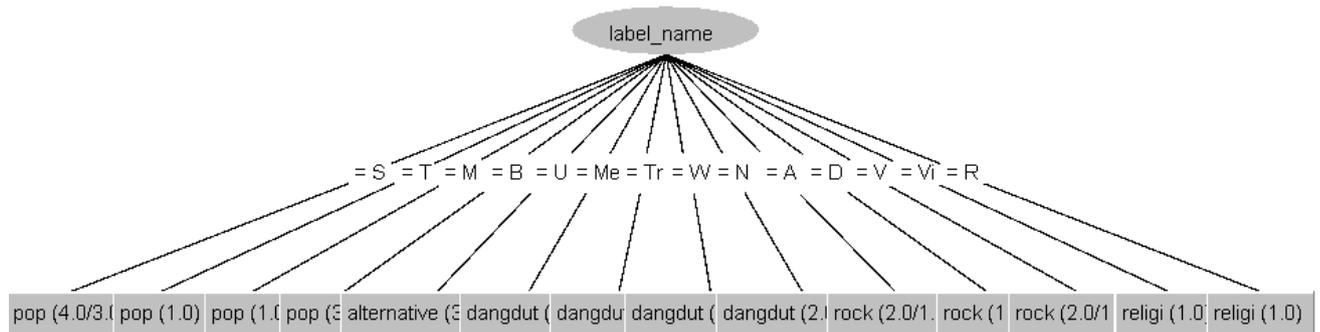
=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
-----
0.882    0.072    0.900     0.882   0.891     0.812   0.969    0.958    Indonesia - Dangdut
0.928    0.118    0.914     0.928   0.921     0.812   0.969    0.978    Indonesia - Pop
Weighted Avg.   0.909    0.099    0.908     0.909   0.908     0.812   0.969    0.969

=== Confusion Matrix ===

      a      b  <-- classified as
378658  50892 |  a = Indonesia - Dangdut
42005  543763 |  b = Indonesia - Pop
    
```

Gambar 5. Hasil pengujian metode algoritma C4.5



Gambar 6. Hasil pengujian *Decision Tree* menggunakan aplikasi *WEKA*

Dari hasil pengujian digunakan tes Cros Validation Algoritma C4.5 menghasilkan nilai Accuracy sebesar 90.8505%. Dengan demikian, dapat disimpulkan prediksi tangga lagu setelah menggunakan metode C4.5 adalah Pop, Alternative, Dangdut Rock dan Religi dengan Label Sony Music, Trinity, Musica Studio.

**5. KESIMPULAN**

Berdasarkan hasil dari pengujian model prediksi pada aplikasi LangitMusik dapat diambil beberapa kesimpulan sebagai berikut :

1. Penerapan *data mining* dengan model algoritma C4.5 dapat digunakan dalam melakukan prediksi tangga lagu pada aplikasi LangitMusik. Berdasarkan hasil penerapan algoritma C4.5 untuk *data testing* menggunakan tabel penilaian *Multi-confussion matrix* menghasilkan *Accuracy* sebesar 90.8505%
2. Dengan adanya model prediksi pada aplikasi LangitMusik dengan algoritma C4.5 maka dapat mempermudah LangitMusik dalam memprediksi tangga lagu.

**6. SARAN**

Penelitian ini dapat dikembangkan lagi dengan atribut, metode dan studi kasus yang berbeda untuk memperoleh hasil yang lebih baik dan relevan, dan juga perlu diadakan pelatihan yang dapat meningkatkan pengetahuan dalam penggunaan algoritma C4.5 pada software *WEKA* sehingga dapat mempermudah dan dapat mempersingkat waktu pengerjaan. Serta, kedepannya perlu mempertimbangkan pada atribut baru tersebut

dalam pendataan arsip. Kemudian untuk penelitian selanjutnya diharapkan menggunakan regresi agar mencapai nilai akurasi yang lebih tinggi.

#### DAFTAR PUSTAKA

- [1] APJII. (2018). Hasil Survei Penetrasi dan Perilaku Pengguna Internet Indonesia 2018. <https://apjii.or.id/survei>
- [2] 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*.
- [3] Emha Taufiq Luthfi. (2009). Penerapan Data Mining Algoritma Asosiasi Untuk Meningkatkan Penjualan. *Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III, 10(1)*, 1–21.
- [4] Shearer, Colin (2000). The CRISP-DM Model: The New Blueprint for DataMining. *JOURNAL of Data Warehouse, Vol. 5, No.4*, p.13-22
- [5] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. Chapman and Hall/CRC (1st ed., Vol. 19). Chapman and Hall/CRC
- [6] Desta Gumilar, Tacbir Hendro Pudjiantoro, Rezki Yuniarti. PREDIKSI KEPOPULERAN LAGU BERDASARKAN TANGGA LAGU BILLBOARD MENGGUNAKAN DECISION TREE DAN K-MEANS.(2017). <https://jurnal.umk.ac.id/index.php/SNA/article/view/1481/1018>
- [7] Solichin, Achamd.(2020)."Model Pembelajaran Decision Tree" | Kecerdasan Artifisial. <https://www.youtube.com/watch?v=KbSLvWaJFxQ>