KOMPARASI ALGORITMA KLASIFIKASI DENGAN MENGGUNAKAN ANACONDA UNTUK MEMPREDIKSI RAMAI PENONTON FILM DI BIOSKOP

Rano Agustino

Program Studi Sistem Informasi, Fakultas Komputer
Universitas Mohammad Husni Thamrin
Jakarta, Indonesia
rano.agustino@gmail.com

Abstract

In the interest of Moviegoers with trendy films, cinemas also play a major role in attracting audiences to watch films they like. But changes that are quite dynamic from audience interest take turns, sometimes it is sometimes not crowded. Thus sometimes the cinema manager experiences an error in placing the film to be aired, so the number of viewers in the cinema is not as expected. From this problem, researchers are interested in analyzing data relating to film audiences in the cinema. By using CART classification, NBC (Naive Bayes Classifier) algorithm, SVM (Support Vector Machine), LR (Logis Text Regression) and LDA (Linear Discriminant Analysis) which will be compared which accuracy is the best for predicting the absence or absence of the audience. Researchers use Anaconda to compare six algorithms and will see the highest results from the Confusion Matrix and ROC Curve

Keywords*: Compare Classification Algorithm, Anaconda for Data Mining, SVM, CART, NBC,SVM, LDA*

I. PENDAHULUAN

Penelitian yang dilakukan oleh Karl Person pada tahun 2015 tentang memprediksi raitng film yang akan di pasarkan, Karl person menggunakan data dari dataset IMDB Movie sebanyak 3376 records, sedangkan algoritma yang digunakan untuk mengukur akurasi ketepatan prediksi menggunakan RF (*Random Forest*) dan SVM (*Support Vector Machine*). Dari hasil perbandingan akurasi yang didapat dengan menggunakan aplikasi Rapid Miner maka menghasilkan nilai RMSE dari kedua model tersebut dengan hasil yang didapat yaitu nilai RF lebih baik dengan dibanding dengan SVM. Dimana nilai RMSE nya yaitu 0.86 (+/0.04).

Masih terkait dengan penelitan yang dilakukan Karl Person, peneliti mencoba mengambil kasus atau permasalahan tentang antusias penonton dalam menonton film. Tetapi ada perbedaan pada penelitian ini, dimana peneliti berfokus kepada jumlah penonton dibioskop saja dan juga cara pengambilan data dan proses dalam pemilihan model algoritma yang akan di uji. Pada penelitian ini penulis berfokus pada cara pengambilan data dan evaluasi dalam pengolahan data untuk memprediksi sepi dan tidak nya pengunjun atau penonton film dibioskop.

Ada beberapa kebijakan yang diambil oleh pihak manajemen bioskop untuk memutuskan menayangkan film dan mengganti film yang sedang show karena dianggap kurang ramai. Ramai atau tidak penonton di tentukan dari jumlah penonton yang akan menonton film tersebut. Standar yang diberikan puhak manajemen untuk menentukan sepi atau tidak nya dihitung dari jumlah penonton < 10 maka dikategorikan sepi.

Dari beberapa kebijakan manajemen tersebut maka peneliti ingin mencoba menganalisa dengan mengklasifikasi sepi atau tidak nya penonton dengan menggunakan beberapa model algoritma klasifikasi sebagai perbandingan untuk menentukan model apa yang akurasi nya lebih baik. Model algoritma yang digunakan. Untuk penelitian ini model algoritma yang digunakan adalah model aloritma klasifikasi diantara nya terdiri dari LR (Logistics Regression), LDA (Linear Discrimi-nant Analysis) SVM (*Support Vector Machine*), K-NN (K-Neighbors Clasification), Decision Trees CART dan NBC (Naive Bayes Classi-fication), dan peneliti menggunakan bahasa pemograman Python .

II. METODE PENELITIAN

Metode penelitian ini mengikuti standard model dari Larose Daniel yaitu CRISP-DM (*Data Mining Cross Industry Standard Process for Data Mining*) [6] yang terdiri dari tahapan pemahaman bisnis, tahapan pemahaman data, tahapan penyiapan data, tahapan pemodelan dan tahapan evaluasi. Bisa dilhat dari gambar 1 Metode Penelitian, dan berikut ini ada lah pemaparan nya

2.1 TAHAPAN PEMAHAMAN BISNIS

Pada tahapan ini terdiri dari latar belakang penelitian, masalah penelitian, tujuan penelitian dan manfaat penelitian yang mana ini sudah diuraikan pada bagian pendahuluan.

2.2 TAHAPAN PEMAHAMAN DATA

Data yang digunakan dalam penelitian ini diperoleh dari database 3 (tiga) lokasi bioskop di Jakarta untuk periode 3 februari 2017 sampai dengan 3 maret 2017. Tiga bioskop tersebut mewakili dari tiga lokasi berikut ini :

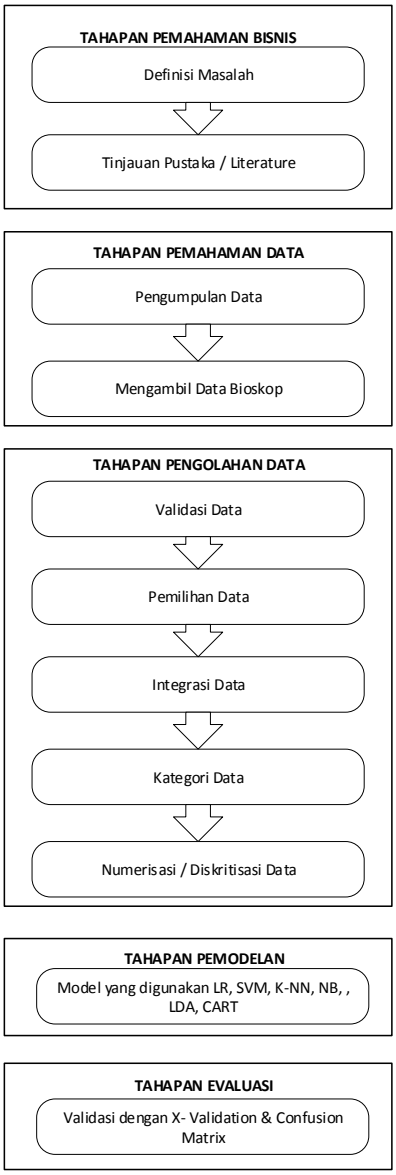
a) Kategori Perumahan : Lokasi bioskop yang berada disekitar perumahan yang berpo-tensi penonton yang datang adalah keluarga seperti anak dan orang tua nya. Untuk perwakilan kategori ini penulis

mengambil data pada lokasi bioskop gading dengan data yang didapat 4586 record data transaksi dan 120 record data film.

b) Kategori Kampus : Lokasi bioskop berada dilingkungan kampus, dimana posisi gedung bioskop berdekatan dengan kampus yang berpotensi penontonnya adalah mahasiswa atau anak remaja. Untuk perwakilan kategori ini penulis mengambil data pada lokasi bioskop Taman Ismail Marzuki dengan data yang didapat adalah 5065 record data transaksi dan 120 record data film

c) Kategori Perkantoran : Lokasi berada disekitar pekantoran yang berpotensi penontonnya adalah kalangan pekerja atau karyawan. Untuk perwakilan kategori ini penulis mengambil data pada lokasi Plaza Indonesia dengan data yang didapat adalah 6034 record data transaksi dan 120 record data film.

Pemilihan tiga bioskop pada tiga jenis lokasi yang berbeda dilakukan untuk melihat karakteristik atau prilaku penonton terhadap jenis film-film yang ditayangkan pada bioskop-bioskop tersebut. Dengan kata lain, peneliti ingin menganalisis jumlah penonton terkait jenis lokasi dan film-film yang ditayangkan pada bioskop di lokasi tersebut.



Gambar 1 Tahapan Penelitian

2.3 TAHAPAN PENGOLAHAN DATA

Data yang digunakan penelitian ini terdiri dari data transaksi dan data film. Data tersebut perlu disiapkan untuk fase selanjutnya, yaitu pembuatan model. Berikut adalah tahapan tahapan yang terdapat dalam penyiapan data [6].

a) Validasi Data

Kualitas data masukan akan menjadi kurang baik jika data tersebut kurang lengkap, tidak konsisten dan tidak rapi [8]. Untuk penelitian ini data yang kurang lengkap dan tidak ada pengaruh nya terhadap klasifikasi maka akan dieliminasi.

b) Pemilihan Data

Beberapa atribut yang dipilih adalah atribut yang diperlukan untuk pengolahan data saja[6],[7],[8], yang mana atribut-atribut tersebut mempengaruhi

hasil klasifikasi dan prediksi. Berikut atribut dari data transaksi yang dipilih dan penjelasannya;

- cinema_id, atribut ini digunakan untuk mewakili lokasi dari bioskop tersebut.
- studio_id, atribut ini digunakan untuk pengelompokan studio disetiap bioskop. Atribut ini juga akan digunakan dalam perhitungan dalam menentukan kategori sepi atau tidak penonton yang dilakukan pada tahap selanjut nya.
- show_id, atribut ini digunakan untuk mengetahui jumlah penonton pada setiap pertunjukan-nya disetiap studio.
- Movie_id, atribut ini sebagai penghubung (Relationship) dengan data film, atribut ini diperlukan agar penggabungan data film dan data transaksi dapat dilakukan.
- Date_Show, atribut ini digunakan untuk atribut hari dan kategori bulan pada fase selanjut nya.

Sedangkan data film yang dipilih dan penjelasannya sebagai berikut:

- Movie_id, atribut ini sebagai penghubung (Relationship) dengan data transaksi, atribut ini diperlukan agar penggabungan data film dan data transaksi dapat dilakukan.
- Category_ID, atribut ini digunakan untuk pembuatan kategori jenis film lokal dan barat.
- Age, atribut ini digunakan untuk pembuatan kategori usia.

c) Integrasi Data

Data Integration adalah proses untuk menggabungkan data dari beberapa sumber [6],[8]. Data transaksi dan data film dengan data yang dipilih pada fase sebelum nya di-integerasikan atau digabungkan menjadi satu dataset untuk pemodelan. Frekuensi Distribusi Data

Ini merupakan susunan data menurut kelas interval tertentu atau menurut kategori tertentu (Hasan, 2001). Pada penelitian ini untuk setiap film yang ditayangkan pada tanggal dan waktu pertunjukan tertentu disuatu bioskop (kelas) dihitung total jumlah penontonnya (Frekuensi).

Dalam proses perjumlahan penonton berikut penjelasan nya:

- Memilih atau mengurutkan dengan mem-filter atribut date_show pada tanggal yang akan dihitung jumlah total penonton lalu dilanjutkan dengan memilih atribut studio_id dan show_id, setelah data dikelompokan akan terlihat jumlah penonton perstudio dan pershow
- Dilanjutkan dengan menjumlahkan total penonton pada pertunjukan selanjutnya dengan cara seperti sebelum-nya maka hasilnya akan terlihat seperti tabel 1.
- Selanjutnya, untuk setiap kelas (pertunjukan film) dilakukan pengkategorisasian sepi atau tidak sepi berdasarkan jumlah penonton-nya. Penulis mendapatkan penjelasan dari pihak managemen bioskop bahwasetiap pertunjukan film (show) yang berjumlah penonton 10 orang atau kurang dikategorikan sepi dan pertunjukan film yang berjumlah penonton 11 orang atau lebih dikategorikan tidak sepi

Tabel 1, Penjumlahan Total Penonton

date_show	Cinema_id	Day	studio_id	show_id	Category_ID	Age	Total Penonton
1/1/2015	JKTTIM	Thursday	2	1	1	15	62
1/1/2015	JKTTIM	Thursday	1	1	1	15	93
1/1/2015	JKTTIM	Thursday	4	1	1	15	32
1/1/2015	JKTTIM	Thursday	3	1	1	15	67
1/1/2015	JKTTIM	Thursday	1	2	1	15	178
1/1/2015	JKTTIM	Thursday	4	2	1	15	65
1/1/2015	JKTTIM	Thursday	3	2	1	15	103
1/1/2015	JKTTIM	Thursday	2	2	1	15	152
1/1/2015	JKTTIM	Thursday	1	3	1	15	207
1/1/2015	JKTTIM	Thursday	4	3	1	15	41
1/1/2015	JKTTIM	Thursday	3	3	1	15	82
1/1/2015	JKTTIM	Thursday	2	3	1	15	143
1/1/2015	JKTTIM	Thursday	1	4	1	15	202
1/1/2015	JKTTIM	Thursday	3	4	1	15	77
1/2/2015	JKTTIM	Friday	1	1	1	15	108
1/2/2015	JKTTIM	Friday	2	1	1	15	64

Tabel 2, Pengkategorisasian pertunjukan film

Kata Lokasi	Day	Show_id	Kat Produksi	KatUsia	Kat Bulan	KatSepi
Perumahan	Thursday	Pertama	Barat	Remaja	Muda	TIDAK
Perumahan	Thursday	Pertama	Barat	Remaja	Muda	TIDAK
Perumahan	Thursday	Pertama	Lokal	Remaja	Muda	TIDAK
Perumahan	Thursday	Pertama	Barat	Remaja	Muda	TIDAK
Perumahan	Thursday	Kedua	Barat	Semua_Umur	Muda	TIDAK
Perumahan	Thursday	Kedua	Lokal	Remaja	Muda	TIDAK

Berikut adalah penjalasan masing-masing atribut pada tabel 3:

- KatLokasi ialah lokasi keberadaan bioskop itu , terdiri dari tiga jenis yaitu perumahan, perkantoran dan kampus.
- Day ialah nama hari
- Show_id ialah waktu pertunjukan yang terdiri dari enam jam waktu tiap harinya, jumlah waktu pertunjukan ini tergantung pada durasi film yang ditayangkan, namun pada umum nya bila waktu durasi film tersebut lebih kurang dari 80 menit maka jumlah waktu pertunjukan adalah 6 kali.
- KatProduksi ialah Kategori Produksi film yang terdiri dari kategori 1 (film barat) dan 2 adalah (film lokal)
- KatUsia ialah kategori usia penonton yang terdiri dari 3 kategori yaitu semua umur, remaja dan dewasa.
- KatBulan ialah kategori bulan terdiri dari 3 kategori yaitumuda dari tanggal 1 sampai dengan 10, sedang dari tanggal 11 sampai dengan 20 dan tua dari tanggal 21 sampai dengan 30 atau 31.
- KatSepi ialah Kategori Sepi yang terdiri dua kategori Ya dan Tidak. Kategori YA jika pe-nonton kurang dari 11 dari setiap pertunjukan.

Setelah tahap ini dilakukan maka didapat dataset yang terdiri dari 1700 record dengan 7 atribut.

d) Numerisasi atau Diskritisasi Data

Untuk menggunakan algoritma Klasifikasi maka dataset diatas perlu didiskritasikan agar isi data menjadi angka.

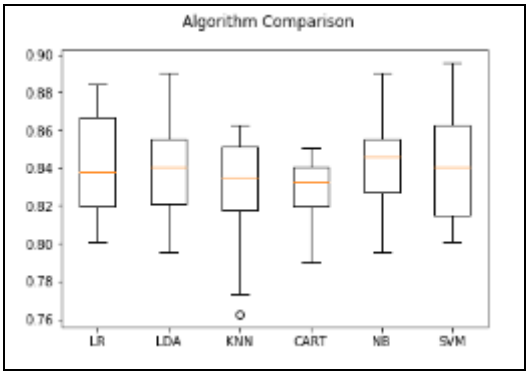
III TAHAPAN PEMODELAN

Pada tahap Pemodelan ini dilakukan pembuatan model klasifikasi dengan menggu-nakan dataset yang telah disiapkan pada tahap sebelumnya. Penelitian ini menggunakan enam algoritma klasifikasi yaitu CART, K-NN, NB, LDA, LR dan SVM, dimana 6 (enam) algoritma tersebut termasuk 10 algoritma terbaik dalam datamining [9].

Dengan menggunakan enam algoritma tersebut dibuatlah empat model klasifikasi dan dilakukan eksperimen untuk mencari model apa yang paling baik dari sisi akurarsi berdasarkan keakuratan dalam pengklasifikasian sepi atau tidak sepinya penonton .

IV PENGUJIAN DAN PEMILIHAN MODEL.

Evaluasi dilakukan dengan metode 10-fold cross validation [7]. Peniliti menggunakan Aplikasi Anaconda Python untuk melakukan komparasi model Alggoritma nya dan hasil yang didapatkan lebih tinggi dengan menggunakan SVM seperti Gambar 2 Hasil Komparasi Algoritma, dan juga pada Gambar 3 Hasil Akurasi Terbaik, dan juga Tabel 5. Hasil Pengujian AUC menjelaskan bahwa SVM mendapat nilai tertinggi yaitu 0.841919.



Gambar 2. Hasil Komparasi Algoritma

Gambar 3 Hasil Akurasi Terbaik

Tabel 5,Hasil Pengujian AUC

Algoritma	CART	SVM	K-NN	NB	LDA	LR
AUC	82	84.19	82	84	83	84.13

Berdasarkan hasil pengujian nilai Accuracy dari model model yang dibuat dengan algoritma CART, K-NN, SVM, LDA dan LR termasuk pada kelas Good classification [7].

V KESIMPULAN DAN SARAN

5.1 Kesimpulan

Penelitian ini berfokus pada penentuan ramai atau sepi nya penonton yang ada di bioskop-bioskop Jakarta. Untuk melihat karakteristik penonton, lokasi bioskop dibagi menjadi tiga kategori, yaitu perumahan, kampus, dan perkan-toran. Dari setiap kategori lokasi ditetapkan

satu bioskop sebagai wakil dan dataset sebanyak 1706 record dengan 7 atribut dianalisis dengan menggunakan 6 (enam) algoritma yang memiliki konsep learning yang berbeda, yaitu SVM, K-NN, NB, LR, LDA dan CART.

Dari hasil evaluasi penulis dengan menggunakan bahasa pemograman python 3.x Anaconda maka menghasilkan nilai yang dapat ditarik kesimpulan bahwa Algoritma SVM mampu memberikan hasil akurasi yang terbaik dibandingkan dengan algoritma lain. Hal ini terbukti dengan hasil akurasi SVM bernilai 84.14%.

5.2 Saran

Penelitian ini dapat dikembangkan untuk memprediksi keterkaitan dalam film-film yang ditonton oleh para pengunjung dengan menggunakan dataset dari luar bioskop, seperti data rating film dari situs web *imdb* atau dari respon masyarakat yang disampaikan melalui media sosial seperti *Twitter* dan *Facebook*.

Daftar Pustaka

Mujain, Saiful, (2015). "*Memahami Pola Menonton Kelas Menengah Muda Urban"*, SMRC, Jakarta.

Persson, (2015). Karl,*PREDICTING MOVIE RATINGS A comparative study on random forests and support vector machines.*"Sweden : University Of Skovde.

Kinnear, Thomas C. and James R. Taylor, (1995).*Marketing Research: An Applied Approach*. "McGraw Hill Text".

Han, Jiawei dan Kamber, Micheline, (2006). *Data Mining : Concept and Techniques Second Edition*, Morgan Kaufmann Publishers.

Patil, S Sherekar, (April 2013)*Performance Analysis Of Naive Bayes and J48 Classification Algorithm for Data Classification*, "International Journal of Computer Science", Vol. 6, No 2.

Larose, D. T, (2005).*Discovering Knowledge In Data An Introduction to Data Mining.* A John Wiley & Sons, Inc., Publication, (2005).

Gorunescu, F., (2011).*Data Mining Concepts, Models and Techniques.* Springer, 1st Edition.

Vercellis, C, (2009).*Business intelligence: Data Mining and Optimization for Decision Making*, "John Wiley & Sons Ltd", Southern Gate, Chichester, *West* Sussex.

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.H., Steinbach, M., Hand, D.J. & Steinberg, D,*Top 10 Algorithms in Data Mining. Survey Paper*. "DOI 10.1007/s10115-007-0114-2", Springer-Verlag, London, (2007).

Demsar, Jane, *Statistical comparisons of classifier over multiple datasets."*Journal of Machine learning.

Brook , John.*SUS - A quick and dirty usability scale*. "Redhatch Consulting Ltd"., United King-dom, 1980