KLASIFIKASI PENJUALAN APLIKASI ANDROID MENGGUNAKAN ALGORITMA C4.5

Alfian Faiz Izzulhaq[1], Sulastri[2]
Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Stikubank
e-mail: [1]alfianfaiz99@gmail.com, [2]sulastri@edu.unisbank.ac.id

Abstrak

Hadirnya smartphone memberi dampak sangat besar pada semua orang. Pada zaman sekarang ini smartphone bisa digunakan untuk mengerjakan tugas seperti surfing di internet, mengerjakan dokumen, mengedit, sesuatu dan sebagainya dengan berbagai aplikasi yang banyak tersedia di internet. Salah satu website penyedia aplikasi adalah http://codecanyon.net. Dalam website tersebut tersedia aplikasi android yang sangat bervariasi dan pengguna harus mengetahui aplikasi yang termasuk laris dan tidak laris dari berbagai sudut pandang.

Tujuan dari penelitian ini adalah untuk melakukan proses klasifikasi aplikasi android yang termasuk dalam kategori laris dan tidak laris. Data yang digunakan berjumlah 300 aplikasi android secara acak. Pada penelitian ini menggunakan metode decision tree algoritma C4.5. Dari tiga kali percobaan klasifikasi menghasilkan nilai akurasi yang berbeda-beda. Nilai akurasi tertinggi didapatkan dari percobaan dengan data training sebanyak 210 dan data testing sebanyak 90 yaitu menunjukkan tingkat akurasi sebesar 73,3%. Rating merupakan atribut yang paling mempengaruhi aplikasi yang termasuk laris atau tidak laris.

Kata kunci: *Aplikasi Android, Analisis Data Mining, Klasifikasi, Algoritma C4.5.*

1. PENDAHULUAN

Hadirnya *smartphone* memberi dampak sangat besar pada kebiasaan pengguna. Smartphone era dulu hanya digunakan sebagai alat komunikasi antar pengguna, sedangkan era sekarang bisa digunakan untuk mengerjakan tugas seperti surfing di internet, mengerjakan dokumen, mengedit sesuatu, dan sebagainya. Fitur-fitur yang tersedia dalam aplikasi android bervariasi, sehingga pengguna bebas memilih aplikasi yang diinginkan. Pada saat ini, banyak website-website penyedia aplikasi dan yang diambil peneliti adalah website *codecanyon.net*.

Situs codecanyon.net merupakan website penjualan aplikasi mulai dari aplikasi komputer, smartphone, framework, source code program, dan lain-lain. Situs tersebut sebagai wadah bagi developer untuk menjual aplikasi seperti aplikasi perkantoran, multimedia, edukasi, permainan, dan lain-lain. Aplikasi yang dijual memiliki fitur, harga, versi, dan review yang bervariasi. Dengan adanya informasi tersebut, dapat dimanfaatkan untuk menggali sebuah informasi. Maka dari itu, dibutuhkan proses analisis yang tepat sehingga informasi berharga yang dihasilkan dapat membantu pihak-pihak dalam lingkup website codecanyon.net untuk lebih memaksimalkan penjualan selanjutnya. Permasalahannya adalah bagaimana cara mengklasifikasikan penjualan aplikasi android pada website *http://codecanyon.net* berdasarkan klasifikasi laku dan tidak laku.

Data mining merupakan proses menemukan informasi penting yang berguna secara otomatis dalam repositori data besar(Tan, dkk, 2006). Algoritma yang digunakan untuk proses ini adalah algoritma C4.5. Menurut Suntoro, (2019) Algoritma C4.5 merupakan algoritma klasifikasi data yang menggunakan perhitungan entropy dan gain ratio untuk pemilihan atribut menjadi node dan output yang dihasilkan dari algoritma C4.5 berupa pohon keputusan *(decision tree)*. Diharapkan dengan analisa ini akan mengetahui aplikasi yang laku dan tidak laku. Dari latar belakang di atas, akan dibuat analisa klasifikasi penjualan aplikasi dengan data mining yang berjudul "Klasifikasi Penjualan Aplikasi Android menggunakan Algoritma C4.5".

2. TINJAUAN PUSTAKA

Menurut penelitian Faradillah (2013) dari STIMIK Budi Darma Medan menjelaskan bahwa memprediksi kebutuhan pasar sangat sulit dimana itu merupakan permasalahan yang dihadapi perusahaan distribusi. Untuk menemukan kebutuhan pasar dibutuhkan pengenalan karakteristik pelanggan. Maka dari itu, peneliti mencoba memprediksi kebutuhan customer. Data yang diolah merupakan data karakteristik transaksi pelanggan PD. Cipta Sari Mandiri Tanjung Morawa pada masa lalu agar bisa menjadi acuan untuk perusahaan meningkatkan hasil produknya. Atribut yang dipakai merupakan produk yang telah dibeli pelanggan. Dari percobaan data sejumlah 24 orang terdapat pelanggan tetap sejumlah 17 orang dan pelanggan baru sejumlah 7 orang dengan rincian produk Tb Saguku dibeli oleh pelanggan tetap sehingga pihak gudang menyiapkan stok lebih untuk ke depannya. Lalu ada produk MSG yang tidak dibeli pelanggan baru maka dari pihak marketing harus meningkatkan penawaran dan lebih mempromosikan produk tersebut agar lebih menarik perhatian customer.

Penelitian yang dilakukan oleh Sundari (2019) dari Universitas Harapan Medan menjelaskan tentang permasalahan dalam pengambilan keputusan berupa prediksi pembelian sepeda motor yang merupakan tugas dari manager CV Berlian Bintang. Ada banyak faktor yang dilihat dalam memprediksi pembelian sepeda motor dan salah satu yang paling penting adalah jenis sepeda motor. Proses data mining dilakukan dengan teknik klasifikasi

menggunakan algoritma C4.5. Dari hasil analisa yang didapat, faktor desain mem iliki nilai yang paling tinggi yang mempengaruhi pembelian sepeda motor.

Menurut penelitian Yahya (2019) dari Universitas Stikubank Semarang menjelaskan tentang permasalahan mahasiswa baru yang tidak registrasi ulang pada saat sudah dinyatakan diterima univeristas. Maka dari itu, peneliti mencoba memprediksi calon mahasiswa baru dengan melakukan analisis data dan menemukan pola-pola penting dalam data registrasi menggunakan data mining. Analisa dilakukan dengan teknik klasifikasi menggunakan algoritma C4.5 dan Naive Bayes. Proses data mining diawali dengan membuat model klasifikasi dan menggunakan data training sebanyak 1866 record. Dari tiga kali pengujian data mahasiswa baru, algoritma C4.5 memiliki rata-rata tingkat akurasi yang lebih baik yaitu sebesar 87,5% sedangkan Naive Bayes sebesar 86,6%.

3. METODE PENELITIAN

3.1. Metode Analisis Data

Analisis data dilakukan berdasarkan KDD (Knowledge Discovery in Database) :

a. Data Selection
Sekumpulan data operasional diseleksi dan dikumpulan terlebih dahulu sebelum masuk ke tahap penggalian informasi. Data diambil secara manual dengan mencari atribut dari tiap detail aplikasi yaitu kategori, versi minimum, rating, harga dan jumlah terjual. Data diambil secara acak dari kategori "*android full application"* sebanyak 300 data. Contoh data awal bisa dilihat pada tabel 1.

Tabel 1. Sampel Data Awal rincian aplikasi *android ful application.*

Kategori	Rating	Versi Minimum	Harga Dollar	Terjual	Laris
Multimedia	5	5.0	34	454	Ya
Multimedia	4,86	4.0	35	348	Ya
Multimedia	4,77	4,3	20	290	Ya
Developer Tools	5	4,4	69	127	Ya
Social Media	4,5	5.0	22	99	Tidak

b. Data Cleaning
Data cleaning merupakan proses memperbaiki kesalahan pada data seperti menghapus data duplikasi, dan data yang inkonsisten. Berhubung data yang diseleksi peneliti secara manual maka proses cleaning dikerjakan sekaligus diproses pengumpulan data.

c. Data Transformation
Data penjualan aplikasi akan ditransformasi ke dalam skema yang sesuai untuk selanjutnya diproses ke data mining. Maka dari itu data pada atribut dikelompokkan ulang agar lebih rapi. Hasil dari proses transformasi bisa dilihat pada tabel 2.

Tabel 2. Tabel data rincian aplikasi setelah proses transformasi.

Kategori	Versi Minimum	Rating	Harga	Laris
Multimedia	Lollipop	Sangat Baik	Murah	Ya
Multimedia	Ice Cream Sandwich	Sangat Baik	Murah	Ya
Multimedia	Jelly Bean	Sangat Baik	Murah Sekali	Ya
Developer Tools	Kitkat	Sangat Baik	Mahal	Ya
Social Media	Lollipop	Sangat Baik	Murah	Tidak

d. Data Mining
Data mining yang akan dikerjakan adalah metode klasifikasi menggunakan algoritma C4.5 dengan tools RStudio yang nantinya akan menghasilkan rule klasifikasi dalam memprediksi penjualan aplikasi android. Dibawah ini merupakan flowchart proses data mining dengan algoritma C4.5.

e. Evaluation
Proses ini bertujuan untuk menampilkan informasi maupun pola-pola yang telah didapatkan dari proses mining dalam bentuk yang mudah dimengerti oleh orang lain. Pada tahap ini didapatkan hasil prediksi dari algoritma C4.5. Selain prediksi juga menghasilkan tingkat akurasi yang didapatkan dari algoritma C4.5.

3.2. Algoritma C4.5

Algoritma C4.5 merupakan algoritma klasifikasi data yang menggunakan perhitungan entropy dan gain ratio untuk pemilihan atribut menjadi node dan output yang dihasilkan dari algoritma C4.5 berupa pohon keputusan *(decision tree)*(Suntoro, 2019). Algoritma C4.5 memiliki rumus perhitungan sebagai berikut :

a. Menghitung nilai Entropy

$$Entropy\ (S) = \sum_{i=1}^{n} - P_i\ x\ log_2 P_i \quad (1)$$

Keterangan :
S : Himpunan Kasus

Pi : Proporsi dari Si terhadap S
n : Jumlah partisi S

b. Menghitung nilai dari Information Gain

$$Gain(S,A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} x\ Entropy\ (Si) \quad (2)$$

Keterangan :
S : Himpunan kasus
A : Atribut
|S| : Jumlah kasus dalam S
|Si| : Jumlah kasus pada partisi ke-i
n : Jumlah partisi atribut A

c. Menghitung nilai Split Info untuk atribut-atribut yang ada,

$$Split\ Info\ (S,A) = -\sum_{i=1}^{n} \frac{Si}{S} log_2 \frac{Si}{S} \quad (3)$$

Keterangan :
S : Himpunan kasus
A : Atribut
|Si| : Jumlah kasus pada partisi ke-i

d. Mengitung nilai dari Gain Ratio untuk atribut-atribut yang ada,
Gain Ratio (S,A) = (Gain(S,A))/(Split Info(S,A)) (4)
Keterangan :
S : Himpunan kasus
A : Atribut
Gain(S,A) : Info gain dari atribut A
Split Info(S,A) : Split info dari atribut A

e. Atribut yang dipilih menjadi akar (root) merupakan atribut yang memiliki nilai Gain Ratio tertinggi sedangkan atribut yang mempunyai nilai Gain Ratio lebih rendah dari akar akan dijadikan sebagai cabang.
f. Menghitung lagi nilai Gain Ratio dari setiap atribut tidak mengikutsertakan atribut yang telah dipilih menjadi akar (root) pada tahap perhitungan sebelumnya.
g. Atribut yang mempunyai nilai Gain Ratio tertinggi dipilih menjadi cabang pada pembentukan Node 1.1.
h. Melakukan pembentukan cabang dengan mengulangi langkah ke-f dan ke-g sampai dihasilkan nilai Gain = 0 untuk masing-masing atribut yang tersisa.

Setelah melakukan perhitungan untuk data training maka langkah selanjutnya adalah mengevaluasi model klasifikasi pada data uji untuk mengukur akurasi pada model yang telah dibuat. Salah satu cara untuk mengukur akurasi adalah dengan menggunakan Confusion Matrix.

4. HASIL DAN PEMBAHASAN

Proses data mining dilakukan menggunakan aplikasi RStudio. Proses ini diawali dengan membuat model *decision tree*. Pemodelan dibuat menggunakan data training sebanyak 210 data. Rule yang dihasilkan bisa dilihat pada gambar 1. .

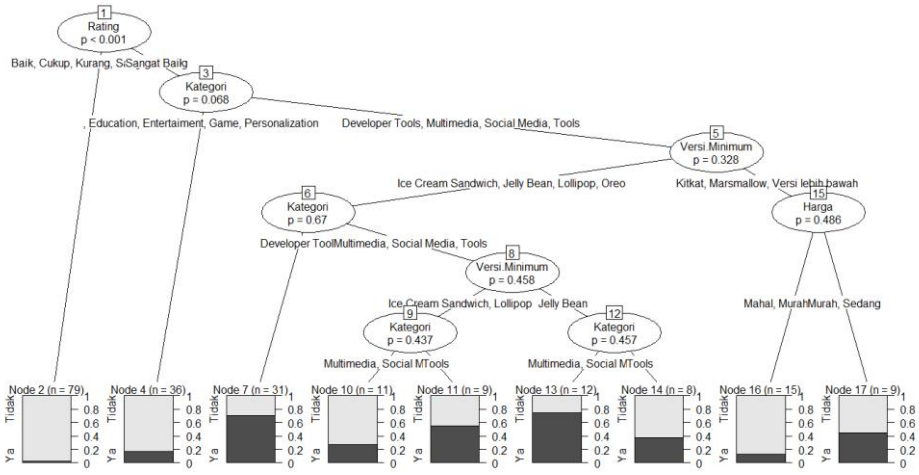
```
Model formula:
Laris ~ Kategori + Versi.Minimum + Rating + Harga

Fitted party:
[1] root
|   [2] Rating in Baik, Cukup, Kurang, Sangat Kurang: Tidak (n = 79, err = 2.5%)
|   [3] Rating in Sangat Baik
|   |   [4] Kategori in , Education, Entertaiment, Game, Personalization: Tidak (n = 36, err = 16.7%)
|   |   [5] Kategori in Developer Tools, Multimedia, Social Media, Tools
|   |   |   [6] Versi.Minimum in Ice Cream Sandwich, Jelly Bean, Lollipop, Oreo
|   |   |   |   [7] Kategori in Developer Tools: Ya (n = 31, err = 29.0%)
|   |   |   |   [8] Kategori in Multimedia, Social Media, Tools
|   |   |   |   |   [9] Versi.Minimum in Ice Cream Sandwich, Lollipop
|   |   |   |   |   |   [10] Kategori in Multimedia, Social Media: Tidak (n = 11, err = 27.3%)
|   |   |   |   |   |   [11] Kategori in Tools: Ya (n = 9, err = 44.4%)
|   |   |   |   |   [12] Versi.Minimum in Jelly Bean
|   |   |   |   |   |   [13] Kategori in Multimedia, Social Media: Ya (n = 12, err = 25.0%)
|   |   |   |   |   |   [14] Kategori in Tools: Tidak (n = 8, err = 37.5%)
|   |   |   [15] Versi.Minimum in Kitkat, Marsmallow, Versi lebih bawah
|   |   |   |   [16] Harga in Mahal, Murah Sekali: Tidak (n = 15, err = 13.3%)
|   |   |   |   [17] Harga in Murah, Sedang: Tidak (n = 9, err = 44.4%)

Number of inner nodes:    8
Number of terminal nodes: 9
```

Gambar 1. *Rule* C4.5 untuk percobaan pertama 210 data.

Dari program yang telah dijalankan menghasilkan aturan pohon keputusan seperti pada gambar diatas. Selanjutnya adalah memvisualkan gambar *decision tree* dari *rule* tersebut. Pohon keputusan bisa dilihat pada gambar 2.



Gambar 2. Pohon keputusan C4.5 percobaan pertama.

Pada gambar diatas menunjukkan bahwa model klasifikasi sudah terbentuk. Atribut Rating menjadi root node yang merupakan atribut paling berpengaruh dalam keputusan aplikasi yang termasuk laris atau tidak laris. Lalu diikuti 8 jumlah internal nodes dan 9 jumlah leaf node.

Tahap selanjutnya adalah menguji model dengan menggunakan data testing sebanyak 90 data. Pengujian ini dilakukan dengan confusion matrix sehingga akan ditemukan nilai akurasi dan nilai kesalahan prediksi. Hasil dari pengujian bisa dilihat pada gambar 3.

```
> hasilconfmatrix
Confusion Matrix and Statistics

          aktual
prediksi Tidak Ya
   Tidak    57 15
   Ya        9  9

               Accuracy : 0.7333
                 95% CI : (0.6297, 0.8211)
    No Information Rate : 0.7333
    P-Value [Acc > NIR] : 0.5547

                  Kappa : 0.2593

 Mcnemar's Test P-Value : 0.3074

            Sensitivity : 0.8636
            Specificity : 0.3750
         Pos Pred Value : 0.7917
         Neg Pred Value : 0.5000
             Prevalence : 0.7333
         Detection Rate : 0.6333
   Detection Prevalence : 0.8000
      Balanced Accuracy : 0.6193

       'Positive' Class : Tidak
```

Gambar 3. Hasil Akurasi *data testing* percobaan pertama.

Pada gambar di atas menunjukkan hasil dari uji coba data testing yang menghasilkan prediksi dan akurasi. Prediksi yang dihasilkan jumlah *class* "tidak" yang diprediksi positif sebanyak 57 data, *class* "ya" yang diprediksi positif sebanyak 9 data, *class* "tidak" yang diprediksi negatif sebanyak, 15 data, dan *class* "ya" yang diprediksi negatif sebanyak 19 data. Lalu untuk nilai akurasi yang dihasilkan sebesar 0,7333 atau bisa dibulatkan menjadi 73,3% dan kesalahan prediksi sebesar 26,7%. Nilai kesalahan prediksi didapatkan dari sisa akurasi sampai angka 100%. Dari rule yang telah terbentuk kemudian diimplementasikan langsung ke data testing, dan diperoleh hasil seperti pada tabel 3 :

Tabel 3. Pengujian *rule* terhadap *data testing* percobaan pertama.

Sample Data Testing					
No	**Kategori**	**Versi Minimum**	**Rating**	**Harga**	**Laris**
1.	Dev. Tools	Kitkat	Sangat Baik	Murah Sekali	Tidak
2.	Dev. Tools	Lollipop	Sangat Baik	Mahal	Ya
3.	Education	Kitkat	Cukup	Murah	Tidak
Sample Rule Decision Tree					
1.	Jika Rating = Sangat Baik, Kategori = Dev. Tools, Versi Mininum = Kitkat, Harga = Murah Sekali, **Maka Tidak Laris**				
2.	Jika Rating = Sangat Baik, Kategori = Dev. Tools, Versi Minimum = Lollipop, Kategori = Dev. Tools, **Maka Laris**				
3.	Jika Rating = Cukup, **Maka Tidak Laris**				

Dari hasil model *decision tree* dan proses uji coba *data testing* dapat disimpulkan bahwa atribut rating sangat berpengaruh untuk pembagian class laris atau tidak laris. Aplikasi yang memiliki nilai rating baik, cukup, kurang, dan sangat kurang dipastikan termasuk kategori aplikasi yang tidak laris. Sedangkan aplikasi yang memiliki rating sangat baik akan masuk ke *node* selanjutnya yaitu *node* kategori dan masih memiliki kemungkinan untuk laris.

Selain itu telah dilakukan dua kali percobaan terhadap data dengan pembagian data training dan data testing. Nilai yang di uji coba sebagai berikut :

a. Percobaan menggunakan Data training 75% (225 data) dan data testing 25% (75 data)
b. Percobaan menggunakan Data training 80% (240 data) dan data testing 20% (60 data)

Tabel 4. Hasil pengukuran akurasi pengujian dari tiga percobaan.

No	Kegiatan	Root Node	Jumlah Prediksi Benar	Jumlah Prediksi Salah	Aktual		Prediksi		Akurasi %	Error %
					TP	TN	FP	FN		
1.	Percobaan 70/30	Rating	66	24	9	57	9	15	73,3%	26,7%
2.	Percobaan 75/25	Rating	53	22	13	40	15	7	70,7%	29,3%
3.	Percobaan 80/20	Rating	41	19	9	32	12	7	68,3%	31,7%

Pada tabel di atas menunjukkan bahwa Rating selalu menjadi root node, atribut rating merupakan atribut yang paling berpengaruh dalam keputusan aplikasi android yang termasuk laris atau tidak laris. Selain itu, pengujian dengan algoritma C4.5 menghasilkan akurasi yang berbeda-beda pada setiap percobaan. Percobaan pertama memiliki nilai akurasi sebanyak 73,3%, percobaan kedua memiliki nilai akurasi 70,7%, dan percobaan ketiga memiliki akurasi sebanyak 68,3%. Selain itu percobaan yang memiliki nilai akurasi tertinggi yaitu dari percobaan pertama dengan nilai akurasi sebanyak 73,3%. Dari ketiga percobaan yang telah dilakukan dapat diambil rata-rata akurasi yaitu sebanyak 70,7%.

5. KESIMPULAN

Berdasarkan hasil penelitian dan pengujian yang telah dilakukan pada klasifikasi penjualan aplikasi android dengan menggunakan algoritma C4.5, maka dapat diambil kesimpulan sebagai berikut :

Dari hasil pengujian data sebanyak tiga kali percobaan didapatkan hasil akurasi terbaik pada percobaan pertama menggunakan 210 data training dan 90 data testing dengan nilai akurasi sebanyak 73,3%. Dari tiga kali percobaan, atribut rating selalu menjadi atribut yang paling mempengaruhi aplikasi yang termasuk laris atau tidak laris.

Atribut Rating yang sangat baik merupakan atribut yang termasuk laris. Kemudian diikuti atribut kategori yang termasuk laris adalah Dev Tools, Multimedia, Social Media, dan Tools. Lalu atribut versi yang termasuk laris adalah Jelly Bean, Ice Cream Sandwich, Lollipop. Sedangkan atribut harga tidak terlalu berpengaruh dalam klasifikasi laris. Dengan bukti rating yang mempengaruhi kelas laris dan tidak laris maka dapat menjadi informasi bagi developer supaya menciptakan aplikasi yang bermanfaat, ramah bagi konsumen, serta user friendly.

Dengan rata-rata hasil akurasi yang mencapai 70,7% maka algoritma C4.5 cukup akurat untuk melakukan klasifikasi penjualan aplikasi android untuk mengelompokkan aplikasi berdasarkan kategori laris dan tidak laris.

6. SARAN

Berdasarkan hasil penelitian yang telah dikerjakan, saran yang diberikan untuk penelitian selanjutnya sebagai berikut :

Melakukan pengujian klasifikasi menggunakan algoritma selain algoritma C4.5 agar mengetahui algoritma yang memiliki akurasi terbaik. Untuk penelitian kedepan sebaiknya menggunakan atribut yang lebih banyak dan bervariasi agar menghasilkan nilai akurasi yang lebih baik.

DAFTAR PUSTAKA

[1] Faradillah, Sarah. (2013) *Implementasi Data Mining untuk Pengenalan Karakteristik Transaksi Customer dengan Menggunakan Algoritma C4.5*. 3 (5), pp. 63-70.
[2] Sundari, Siti. (2019) *Implementasi Data mining dengan Algoritma C4.5 untuk Memprediksi Pembelian Tipe Sepeda Motor*
[3] Suntoro, Joko. (2019). *Data Mining : Algoritma dan Implementasi dengan Pemrograman PHP*. Jakarta : PT Elex Media Komputindo.
[4] Tan, Pang-Ning, Michael Steinbach, & Vipin Kumar. (2006). *Introduction to Data Mining*. Pearson International Edition. United States: Pearson Education Inc.
[5] Yahya, Norzam. (2019). Perbandingan Kinerja Algoritma C4.5 dan Naive Bayes untuk Prediksi Masuknya Calon Mahasiswa Baru (Studi Kasus : Universitas Stikubank Semarang), *Skripsi*, Program Studi Sistem Informasi FTI Unisbank, Semarang.