KOMPARASI ALGORITMA *C4.5* DENGAN *RANDOM FOREST* UNTUK REKOMENDASI PENJUALAN GAUN *ALIEXPRESS.COM*

Abdurrahman Waid, Deni Arifiant M.Kom

Program Studi Teknik Informatika,Fakultas Teknik, Universitas Muhammadiyah Jember

Jl. Karimata No.49 Jember Kode Pos 68121

Email : Abdurrahmanw58@gmail.com

Abstrak - *Aliexpress* adalah layanan ritel *online* yang berbasis di *china* yang dimiliki oleh *Alibaba* yang diluncurkan pada tahun 2010 oleh *Jack Ma*. *Aliexpress* merupakan situs *web e-commerce* yang memfasilitasi usaha kecil untuk menjual kepada pelanggan diseluruh dunia, salah satu penjualan pada toko online *Aliexpress* yaitu gaun wanita. Pada penelitian ini atribut yang digunakan meliputi *Style, Price, Rating, Size, Season, NeckLine, SleeveLength, waiseline, Material, FabricType, Decoration, Pattern Type.* Pada penelitian sebelumya yang dilakukan oleh *V.H.Valentino* tahun 2018 komparasi Metode *C4.5, Naïve Bayes,* dan *Random Forest* untuk mentukan kelulusan mata kuliah. Hasil yang didapat menggunakan metode *C4.5* sebesar 98.89%, *Naïve Bayes* sebesat 96.67%, Sedangkan nilai akurasi *Random Forest* sebesar 95.56%. Pada penelitian yang akan dilakukan dengan data penjualan gaun *Aliexpress* diharapkan dapat mengertahui algoritma mana yang terbaik antara *C4.5* dan *Random Forest*.

Setelah melakukan pengujian sebanyak 41 kali dengan menggunakan teknik *cross validation*, diketahui hasil terbaik yaitu dengan menggunakan metode *Random Forest* pada K20 percobaan 10 dimana hasil akurasi yang didapat sebesar 72.00% , presisi sebesar 65.00%, sedangkan *C4.5* menghasilkan akurasi terbesar 66.00%, presisi 69.23% pada K10 percobaan 10. Diketahui nilai akurasi dan presisi *Random Forest* lebih unggul dengan selisih 6% .

Kata Kunci - *Komparasi Algoritma C4.5 Dengan Random Forest*

PENDAHULUAN

Aliexpress adalah layanan ritel *online* yang berbasis di *china* yang dimiliki oleh *Alibaba* yang diluncurkan pada tahun 2010 oleh *Jack Ma*. *Aliexpress* merupakan situs web *e-commerce* yang memfasilitasi usaha kecil untuk menjual kepada pelanggan diseluruh dunia, Penjual di *Aliexpress* dapat berupa perusahaan atau perorangan. Pada penelitian sebelumnya yaitu komparasi metode *C4.5, Naïve Bayes,* dan *Random Forest* yang dilakukan oleh *V.H.Valentino* (2018), dalam penelitinya mereka melakukan perbandingan nilai akurasi untuk mentukan kelulusan mata kuliah. Hasil yang didapat ketika menggunakan metode *C4.5* sebesar 98.89%, *Naïve Bayes* sebesar 96.67%, sedangkan nilai akurasi *Random Forest* sebesar 95.56%.

Pada penelitian yang dilakukan, komparasi algoritma klasifikasi *C45* dan *Random Forest* digunakan sebagai algoritma yang membentuk rekomendasi penjualan gaun dalam bentuk pohon keputusan. Hal ini didasarkan atas keunggulan algoritma *C45* yang mampu memangkas struktur hierarki pohon keputusan terhadap parameter yang dimiliki dalam proses klasifikasi sehingga memudahkan dalam implementasi pengambilan keputusan. *Random Forest* merupakan pengembangan dari *Decision Tree*, dimana setiap *Decision Tree* telah dilakukan training menggunakan sampel individu dan setiap atribut dipecah pada *tree* yang dipilih antara atribut subset yang bersifat acak. Dan pada proses klasifikasi, individunya didasarkan pada vote dari suara terbanyak pada kumpulan populasi *Tree*. Berdasarkan uraian diatas penulis mencoba menguji dan membandingkan dua algoritma yaitu *C4.5* dengan *Random Forest* pada data penjulan gaun *Aliexpress.com*.

TINJAUAN PUSTAKA

Toko *Online*

Toko *Online* merupakan tempat pembelian barang dan jasa melalui media *internet*, seorang pembeli dapat melihat terlebih dahulu barang yang akan di belanjakan melalui *web* yang dipromosikan oleh penjual. Toko *Online* adalah salah satu bentuk perdagangan elektronik (*Ecommerce*) barang atau jasa yang di lakukan tanpa bertatap muka secara langsung. *Online shop* memberikan beragam kemudahan bagi konsumennya diantaranya adalah adanya penghematan biaya, barang bisa langsung diantar ke rumah, pembayaran dilakukan secara transfer, dan harga lebih bersaing (Juju & Maya, 2010).

C4.5

Algoritma *C4.5* merupakan Algoritma yang di gunakan untuk membentuk sebuah pohon keputusan *(Decsion Tree)*, Pohon keputusan berguna untuk mengekspolari data, menemukan hubugan tersembunyi. *C4.5* merupakan pengembangan dari *ID3*, beberapa pengembangan yang dilakukan pada *C4.5* antara lain bisa mengatasi missing value, bisa mengatasi continue data, dan pruning. Proses pada pohon keputusan adalah mengubah bentuk data(tabel) menjadi model pohon, mengubah model pohon menjadi rule, dan menyederhanakan rule (Basuki dan Syarif, 2003). Secara umum algoritma *C4.5* untuk mambangun pohon keputusan adalah sebagai berikut :

1. Pilih atribut sebagai akar (*root*).
2. Buat cabang untuk masing-masing nilai.
3. Bagi kasus dalam cabang.
4. Ulangi proses untuk masing-masing cabang sampai semua kasus pada cabang memiliki kelas yang sama.

$$Entropy(S) = \sum_{i=1}^{n} - pi * log_2 pi$$

Dimana :

S : himpunan kasus

n : jumlah partisi S

pi : proporsi Si terhadap S

Untuk memilih atribut akar, didasarkan pada nilai gain tertinggi dari atribut-atribut yang ada. Untuk menghitung gain digunakan rumus seperti yang tertera dalam persamaan berikut :

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

Dimana :

S : himpunan kasus

A : atribut

n : jumlah partisi atribut A

|*Si*| : jumlah kasus pada partisi ke-i

|*S* | : jumlah kasus dalam S

Random Forest

Random Forest adalah pengklasifikasi yang terdiri dari kumpulan pengklasifikasi pohon terstruktur {h(x,Θk), k=1, ...} dimana {Θk} adalah vektor acak terdistribusi yang identik independen dan masing-masing pohon melemparkan unit suara untuk kelas paling populer di input x. *Random Forest* merupakan pengembangan dari Algoritma *C4.5* (*decision tree*) dengan menggunakan beberapa *Decision tree*, dimana setiap *Decision Tree* telah dilakukan *training* data menggunakan sampel individu dan setiap atribut dipecah pada *Tree* yang dipilih antara atribut subset yang bersifat acak. Dan dalam perkembangannya, sejalan dengan bertambahnya dataset, maka *Tree* pun ikut berkembang. Penempatan *tree* yang saling berjauhan membuat apabila terdapat *Tree* disekitar *Tree* x berarti pohon tersebut merupakan perkembangan *Tree* x.

Random Forest merupakan salah satu algoritma klasifikasi dengan tingkat akurasi yang baik. *Random Forest* merupakan sebuah metode *ensemble* yang terdiri dari beberapa pohon keputusan sebagai *classifier*. Kelas yang dihasilkan dari proses klasifikasi ini diambil dari kelas terbanyak yang dihasilkan oleh pohon-pohon keputusan yang ada pada *Random Forest*. Dengan melakukan voting pada pohon-pohon keputusan yang tersedia membuat akurasi dari *Random Forest* meningkat.

Rapid Miner

Rapid Miner merupakan perangakat lunak yang bersifat terbuka (*open source*).*Rapid Miner* adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. *Rapid Miner* menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. *Rapid Miner* memiliki kurang lebih 500 operator data mining, termasuk operator untuk input, output, data *preprocessing* dan visualisasi. *Rapid Miner* merupakan *software* yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang dapat diintegrasikan pada produknya sendiri. *Rapid Miner* ditulis dengan munggunakan bahasa *java* sehingga dapat bekerja di semua sistem operasi.

Rapid Miner menyediakan *GUI* (*Graphic User Interface*) untuk merancang sebuah pipeline analitis. *GUI* ini akan menghasilkan file *XML* (*Extensible Markup Language*) yang mendefenisikan proses analitis

keingginan pengguna untuk diterpkan ke data. File ini kemudian dibaca oleh *Rapid Miner* untuk menjalankan analis secara otomatis.

Confusion Matrix

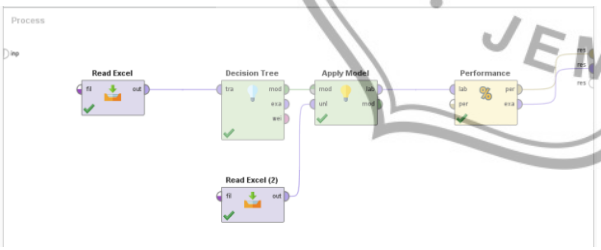
Tahap berikutnya melakukan pengujian terdapat 2 bagian :

1. Akurasi
   Didefinisikan sebagai tingkat kedekatan antara nilai prediksi dengan nilai aktual.
2. Presisi
   Tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem.

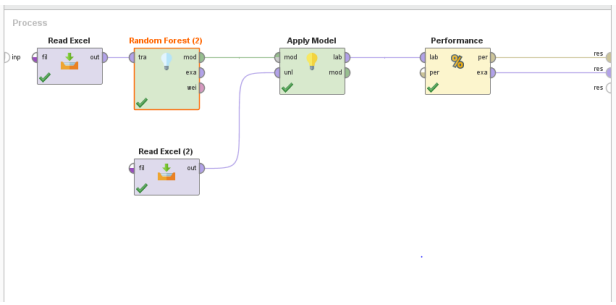
PENGUJIAN

Pengolahan Data Awal

Dalam pengujian ini menggunakan rapid miner dengan cross-validation untuk mendapatkan hasil Akurasi dan Presisi pada setiap algoritma yang diuji menggunakan dataset penjualan *Aliexpress.com.*



Gambar 3.1 Desain Pengujian menentukan akurasi dan presisi *C4.5*



Gambar 3.2 Desain Pengujian menentukan akurasi dan presisi *Random Forest*

Pengukuran Penelitian

Berikut adalah hasil semua percobaan yang dilakukan dengan menggunakan *C4.5* dan *Random Forest*, dapat kita lihat pada Tabel dibawah ini:

Tabel 3.1 Akurasi Tertinggi *C4.5* dan *Random Forest*

	Metode	K = 2	K = 4	K = 5	K = 10	K=20	Hasil Tertinggi
Akurasi	*C4.5*	54.20%	58.40%	61.39%	66.00%	64.00%	66.00%
	Random Forest	53.20%	60.80%	68.32%	66.00%	72.00%	72.00%
Presisi	*C4.5*	41.25%	68.25%	66.67%	69.23%	84.62%	69.23%
	Random Forest	36.62%	66.25%	75.00	66.67%	65.00%	65.00%

KESIMPULAN

Kesimpulan

Berdasarkan penelitian yang sudah dijelaskan pada hasil dan pembahasan sebelumnya yaitu menguji algoritma *C4.5* dengan algoritma *Random Forest* untuk menentukan akurasi dan presisi paling tinggi pada data penjualan toko *online Aliexpress.com* dengan melakukan beberapa pengujian data menggunakan atribut, *Style, Price, Rating, Size, Season, NeckLine, SleeveLength, waiseline, Material, FabricType, Decoration, Pattern Type.*

Setelah dilakukan uji skenario 1 sampai 5, dimana jumlah keseluruhan data dihitung sebanyak 41 kali dengan k-fold 2, 4, 5, 10, 20 menggunakan algoritma *C4.5* dan *Random Forest*. Dalam pengujian data diperoleh nilai akurasi dan presisi yang akan diambil yaitu nilai tertinggi. Diperoleh beberapa nilai tertinggi dari perhitungan *C45* yaitu pada skenario K10 iterasi 10 dengan nilai akurasi 66.00%, presisi sebesar 69.23%. sedangkan metode *Random Forest* memiliki nilai akurasi sebesar 72.00% , presisi sebesar 65.00% pada skenario K20 iterasi 10, *Random Forest* lebih unggul dengan selisih nilai akurasi sebesar 6%.

Saran

Berdasarkan penelitian yang telah di lakukan, diajukan saran penelitian sebagai berikut:

1. Dalam penelitian selanjutnya dapat mengembangkan dan menambah data rekomendasi penjualan agar lebih meningkatkan keakuratan kinerja dari metode yang akan digunakan.
2. Untuk mendapatkan hasil akurasi yang lebih maksimal dapat menambahkan metode *Dizcretization* dan teknik *Bagging*.

REFERENSI

Andira, S, C, (2015), "Perilaku Berbelanja *Online* Di Kalangan Mahasiswi Antropologi Universitas Airlangga", Universitas Airlangga.

Asa, S, R, (2019), "Identifikasi Penyaluran Zakat Menggunakan Algoritma *C4.5* (Studi Kasus Di Baznas Kabupaten Agam)", Fakultas Ilmu Komputer, UPI YPTK Padang.

Ansari, H, D, (2018), "Perilaku Belanja *Online* Di Indonesia", Fakultas Ekonomi Universitas Islam Sumatera Utara

Budi, A, I, M, (2015), "Prediksi Lama Studi Mahasiswa Dengan Metode *Random Forest* (Studi Kasus : Stikom Bali)", STMIK STIKOM Bali.

Fayyad, Usama, (1996), "*Advances in Knowledge Discovery and Data Mining*. *MIT Press*".

Frastian, Nahot, (2018); "Komparasi Algoritma Klasifikasi Menentukan Kelulusan Mata Kuliah Pada Universitas": Fakultas Teknik dan Ilmu Komputer Universitas Indraprasta PGRI.

Hinmaniar, T, S, (2018), "Strategi Internasionalisasi *Aliexpress* (*E-Retail Subsidiaris Internasional Alibaba Group*) Di Rusia", Universitas Airlangga.

Ibrahim, A, Ibrahim (2017), "Modeling of the output current of a photovoltaic grid-connected system using random forests technique", Energy Exploration & Exploitation.

Juju, D., & Maya, M, (2010). "Cara Mudah Buka Toko *Online* dengan *Wordpress+WP E-Commerce*", Yogyakarta: Andi Offet.

Kusnawi, (2007). "Pengantar Solusi Data Mining", STMIK AMIKOM Yogyakarta.


UNIVERSITAS MUHAMMADIYAH
JEMBER