Jurnal Integrasi
Vol. 11 No. 2, October 2019, 140-145
e-ISSN: 2548-9828

Article History
Received September 2019
Accepted October 2019

ANALISIS PERBANDINGAN ALGORITMA NAIVE BAYES, J48,DAN RANDOM FOREST TREE DALAM PENINGKATAN LOYALITAS PELANGGAN UMKM DENGAN VOUCHER BELANJA

Maya Cendana*, Silvester Dian Handy Permana*

∗Universitas Trilogi
Program Studi Teknok Informatika
Jalan TMP Kalibata No. 1, Jakarta Selatan 12760, Indonesia
E-mail: mcendana@trilogi.ac.id, handy@trilogi.ac.id

Abstrak

Teknologi informasi sudah dimanfaatkan sejak lama dalam peningkatan profit dalam usaha UMKM. Banyak masyarakat yang memiliki bisnis UMKM menggunakan toko online untuk mempromosikan bisnisnya. Untuk dapat menarik pelanggan yang lama agar berbelanja kembali ke toko online, salah satunya dengan memberikan voucher belanja. Voucher belanja diberikan untuk pelanggan lama yang mempunyai potensial untuk berbelanja kembali ke toko online. Dalam menentukan pelanggan mana yang tepat dibutuhkan algoritma penambangan data untuk mencari informasi yang tepat di mana pelanggan tersebut dapat berbelanja kembali. Namun kesalahan memilih algoritma dapat mengakibatkan tidak optimalnya pendapatan yang diproyeksikan. Dalam penelitian ini akan menganalisis dan membandingkan algoritma Naive Bayes, J48, dan Random Forest Tree untuk studi kasus toko online. Penelitian ini melibatkan 10 kriteria yang akan digunakan untuk menjadi bahan dalam pengolahan data. Dari hasil penelitian ini didapatkan random forest tree adalah algoritma terbaik untuk menentukan potensial dari pelanggan toko online, yaitu dengan tingkat akurasi sebesar 99,38%, diikuti oleh algoritma J48 dengan tingkat akurasi 81,85%, dan Naïve Bayes sebesar 80,17%. Hasil penelitian ini digunakan untuk membantu proses pengambilan keputusan pemberian voucher belanja kepada pelanggan agar bisnis UMKM dapat berjalan dan mendapatkan keuntungan yang optimal.

Kata kunci: Analisis Perbandingan, Perbandingan Algoritma, Naive Bayes, J48, Random Forest Tree

Abstract

Information technology has been used for a long time for MSME businesses. Many people who have MSME businesses use online stores to promote their businesses. To be able to attract old customers to shop back to the online store, one of them is by giving a shopping voucher. Shopping vouchers are given to existing customers who have the potential to shop back to online stores. In determining which customer is the right data mining algorithm is needed to find the right information where the customer can shop again? But the error of choosing an algorithm can result in not being optimal in the projected income. In this study, we will analyze and compare the Naive Bayes, J48, and Random Forest Tree algorithms for case studies of online stores. This study involved 10 criteria that would be used in data processing. From the results of this study, a random forest tree is the best algorithm to determine the potential of online store customers with an accuracy rate of 99.38%, followed by the J48 algorithm with an accuracy rate of 81.85%, and Naïve Bayes (80.17%). The results of this study are used to help the decision-making process of giving shopping vouchers to customers so that MSME businesses can run and get optimal profits

Keywords: Comparation Analysis, Algorithm Comparation, Naive Bayes, J48, Random Forest Tree

1. Introduction

Perkembangan teknologi informasi kini sudah masuk pada ranah bisnis usaha mikro, kecil dan menengah (UMKM). Banyak masyarakat Indonesia memanfaatkan teknologi informasi berupa Website, toko online, maupun sosial media untuk mengembangkan bisnisnya. Mereka semua berlomba-lomba untuk memenangkan pasar bisnis sesuai dengan bidangnya. Banyak cara yang telah dilakukan untuk mengoptimalkan pendapatan mereka. Tidak jarang dari mereka menggunakan jasa teknologi informasi untuk mengembangkan bisnisnya secara eksternal. Biasanya jasa teknologi informasi digunakan untuk menambah jangkauan ke masyarakat luas. Hal ini biasanya menggunakan promosi sosial media seperti Facebook dan Instagram dimana banyak sekali jasa yang menawarkan jumlah *followers* agar toko onlinenya banyak dikunjungi oleh calon pelanggan [1]

Selain menambah jangkauan, banyak pula bisnis UMKM yang tidak memanfaatkan informasi dari pelanggan sebelumnya. Banyak data berupa transaksi pembayaran dari pelanggan tidak dikelola secara baik oleh pegiat bisnis UMKM. Padahal, dari data transaksi pembelian yang dilakukan oleh pelanggan didapatkan banyak sekali informasi yang dapat digunakan untuk meningkatkan bisnis UMKM. Data yang dapat diolah dari transaksi yang sudah dilakukan oleh pelanggan berupa manajemen stok barang, manajemen supplier, atau bahkan memprediksi pembelian pelanggan. Dari hal tersebut yang jarang dilakukan oleh pegiat bisnis UMKM adalah memprediksi kemungkinan pelanggan untuk membeli lagi barang yang dijual. Agar dapat meningkatkan minat pembelian dari pelanggan yang sudah melakukan transaksi sebelumnya, para UMKM dapat memberikan *voucher* belanja kepada pelanggan lama. Hal ini digunakan untuk menarik minat pelanggan yang sudah ada dengan mengoptimalkan potensi keuntungan yang dapat diterima oleh UMKM [2]

Pemberian voucher belanja kepada pelanggan yang sudah membeli sebelumnya harus dipilih dengan bijaksana. Proses pemberian *voucher* belanja tersebut tidak dapat dilakukan dengan sembarangan karena dapat mengurangi potensi keuntungan yang dapat dicapai oleh UMKM. Dalam proses penentuan pemberian *voucher* tersebut harus dilakukan dengan teknik penambangan data. Data pelanggan berupa pembelian barang terakhir, usia, alamat domisili, jenis kelamin, dsb merupakan hal yang perlu dipertimbangkan untuk menentukan pemberian *voucher* belanja [3]

Dalam proses pengolahan data tersebut mempunyai banyak algoritma yang digunakan. Namun, kesalahan dalam menentukan algoritma penambangan data tersebut dapat mengurangi potensi keuntungan bisnis UMKM yang dapat diraih. Penelitian ini tentunya ingin mengoptimalkan keuntungan yang dapat diraih oleh UMKM dengan data pelanggan yang sudah ada sebelumnya. Penilaian ini menganalisis algoritma Naive Bayes, Algoritma J48, dan Algoritma Random Forest Tree. Ketiga algoritma tersebut merupakan algoritma penambangan data yang cocok untuk memprediksi bisnis dengan memanfaatkan data yang sudah ada dan nantinya harus dianalisis dan diuji secara tepat dan sistematis untuk memproyeksikan pelanggan mana saja yang akan melakukan transaksi lagi di dalam bisnis UMKM.

Penelitian ini diharapkan dapat membantu masyarakat yang memiliki bisnis UMKM agar mendapatkan keuntungan yang optimal. Keuntungan bisnis yang optimal dapat diraih dengan data transaksi yang diambil dan menghasilkan sebuah informasi yang penting. Diharapkan penelitian ini dapat menjadi contoh atau model untuk mengembangkan bisnis di sektor UMKM.

2. Tinjauan Pustaka

Penelitian yang dilakukan oleh [4] menggunakan *machine learning ensembles* untuk prediksi longsoran dalam membantu kinerja pemerintah. Ensembles ini disusun atas kombinasi antara algoritma J48, Adaboost, Bagging, dan Rotation Forest. Kombinasi algoritma berbasis data mining ini menunjukkan dan mendapatkan model performa tinggi untuk memprediksi kelongsoran tanah. Distrik Guangchang (provinsi Jiangxi, Cina) terpilih sebagai studi kasus pada penelitian ini. Penelitian ini meneliti 237 lokasi longsor; lokasi longsor kemudian secara acak dibagi menjadi rasio 70/30 untuk pelatihan dan validasi model. Hal yang dihitung dalam penelitian ini adalah faktor seperti kemiringan, aspek, ketinggian, indeks kelembaban topografi (TWI), kekuatan aliran indeks (SPI), indeks transportasi sedimen (STI), kelengkungan rencana, kelengkungan profil, litologi, jarak ke sesar, jarak ke sungai, jarak ke jalan, penggunaan lahan, indeks vegetasi perbedaan normal (NDVI), dan curah hujan. Hasil dari penelitian ini menunjukkan bahwa JDT dengan Rotation Forest adalah model optimal terbaik dan dapat dipertimbangkan sebagai metode yang menjanjikan untuk pemetaan kerentanan longsor dalam kasus serupa untuk akurasi yang lebih baik. Penelitian ini memberikan informasi mengenai Algoritma Rotation Forest sebagai salah satu algoritma yang terbaik untuk memprediksi kelongsoran tanah.

Penelitian yang dilakukan oleh [5] mengungkapkan bahwa data mining dimanfaatkan untuk melakukan proses analisis data dalam rangka menemukan informasi terstruktur dari data yang dikumpulkan, dan akan membantu dalam pengambilan keputusan. Penelitian ini dilakukan dengan mencari sejumlah *frequent itemset*, lalu membentuk aturan-aturan asosiasi (*association rules*). Penelitian ini

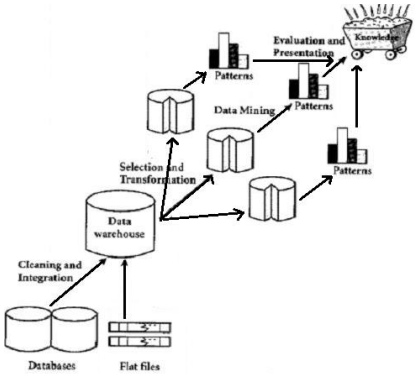
menggunakan Algoritma Apriori dan algoritma *frequent pattern growth* (FP-growth) untuk menemukan sejumlah *frequent itemset* dari data-data transaksi yang tersimpan dalam basisdata. Penelitian ini menggunakan kedua algoritma tersebut untuk membentuk suatu kaitan dengan informasi penjualan buku di PT. Gramedia, sehingga dapat digunakan sebagai pertimbangan pihak manajemen dalam membuat strategi penjualan kedepannya secara efektif. Penelitian ini dapat menjadi salah satu alternatif dalam memperoleh keputusan.

Penelitian dengan judul The Impact of Churn on Client Value in Health Insurance, Evaluation Using a Random Forest under Random Censoring dilakukan oleh [6]. Penelitian ini mengukur dampak Churn pada nilai klien dalam asuransi kesehatan menggunakan evaluasi Random Forest. Nilai yang diukur adalah manfaat yang didapatkan oleh pelanggan dalam memperoleh manfaat asuransi dimana semakin lama pemegang polis mempertahankan kontrak mereka, semakin banyak keuntungan yang ada untuk perusahaan. Oleh karena itu, dibutuhkan prediksi waktu dimana pemegang polis potensial akan menyerahkan kontrak mereka karena hal itu sangat penting untuk mengoptimalkan keuntungan. Penelitian ini menggunakan model random forest untuk mengatasi masalah ini. Model ini dirancang untuk mengimbangi dampak penyensoran acak. Penelitian ini menunjukkan bahwa pendekatan ini sangat kompetitif dalam hal kesalahan kuadrat dalam mengatasi masalah yang diberikan.

Penelitian yang dilakukan oleh [7] mengungkapkan bahwa penelitiannya dapat memperkirakan kekuatan beton yang nantinya akan dihasilkan. Kekuatan beton yang diperkirakan adalah hal yang terpenting didalam membangun infrastruktur seperti gedung atau jalan. Penelitian ini menggunakan penambangan dan pengolahan data dengan metode estimasi guna mengkalkulasi kekuatan beton yang dihasilkan. Penelitian ini memproses data dengan menggunakan algoritma regresi linear dalam perhitungan komponen yang digunakan untuk menghitung kekuatan dari beton. Di dalam pengujiannya, penelitian ini menggunakan Cross Validation dan evaluasinya menggunakan Root Mean Square Error (RMSE) untuk menilai seberapa besar kesalahan dari metode regresi linear. Penelitian ini hanya membangun konsep saja tanpa memiliki kesimpulan studi kasus yang dihasilkan.

3. Metodologi Penelitian

Metode penelitian yang digunakan pada penelitian ini memodifikasi tahapan data mining yang dikemukakan oleh [8] dan ditunjukkan pada Gambar 1.



Gambar 1: Metodologi Penelitian

Alur Metode Penelitian yang digambarkan pada Gambar 1 memiliki beberapa langkah, yaitu:

1. Pembersihan Data
   Pembersihan data yang dilakukan ini digunakan untuk menyaring data-data yang tidak sempurna atau valid seperti data yang tidak terisi seutuhnya, data yang diisi dengan acak, dan sebagainya. Pembersihan data ini bertujuan untuk mendapatkan data yang baik dan siap diolah. Proses ini berpengaruh pada performasi yang akan dijalankan nantinya untuk mencapai hasil yang terbaik.
2. Integrasi Data
   Integrasi data ini digunakan untuk mengidentifikasi atribut dari setiap entitas yang unik. Atribut yang dilihat seperti atribut nama, tanggal, jenis produk, jenis kelamin, nomor pelanggan dsb. Integrasi data perlu dilakukan secara cermat karena kesalahan pada integrasi data bisa menghasilkan hasil yang menyimpang dan bahkan menyesatkan pengambilan aksi nantinya.
3. Data Mining
   Data mining ini dilakukan untuk melihat setiap proses dimana setiap algoritma yang dipakai akan dijalankan pada proses ini. Proses penambangan data ini memasukkan setiap atribut agar mendapatkan informasi yang terbaik.
4. Evaluasi Pola yang Ditemukan
   Dari setiap algoritma nantinya akan menghasilkan pola tersendiri dalam mengimpretasikan data atau informasi yang didapat dari setiap proses algoritma data mining yang dilakukan pada tahap 3.
5. Presentasi Pola yang Ditemukan untuk Menghasilkan Aksi
   Tahap terakhir dari proses data mining adalah memformulasikan keputusan dari hasil pola yang didapat. Dalam bagian ini juga divisualisasikan objek yang didapat dari pola yang ditemukan. Bagian ini juga menguji terbalik kedalam data tes yang disediakan sebelum digunakan dalam pengambilan

keputusan.

4. Pembahasan

Terdapat 32 atribut dalam penelitian ini, tetapi hanya diambil 10 atribut saja yang digunakan sebagai data train dan data uji yang ditunjukkan pada Tabel 1.

TABEL 1

ATRIBUT PENGUJIAN

Atribut	Tipe	Karak-teristik	Keterangan
salutation	integer	nominal	Saturation, seperti Mr., Ms., Company
domain	integer	nominal	Email provider domain, seperti gmail.com, yahoo.com, hotmail.com dll
newsletter	integer	nominal	Newsletter subscribed, seperti yes dan no
paymenttype	integer	nominal	Payment type, seperti payment on invoice, cash payment, transfer from current account, dan transfer from credit card
deliverytype	integer	nominal	Delivery type, seperti dispatch dan collection
case	integer	ordinal	Value of goods, seperti low dan high
voucher	integer	nominal	Voucher redeemed, seperti yes dan no
gift	integer	nominal	Gift option, seperti yes dan no
shippingcosts	integer	nominal	Shipping cost incurred, seperti yes dan no
weight	integer	cardinal	Shipment weight

Random Forest (Data Train)
- Tahap pengujian dilakukan dengan menggunakan 10 fold cross validation.
- Jumlah instances: 32427
- Jumlah atribut: 32
- Correctly Classified Instances: 32227 (99.3802 %)
- Incorrectly Classified Instances: 201 (0.6198 %)

Random Forest (Data Uji)
- Tahap pengujian dilakukan dengan menggunakan 10 fold cross validation.
- Jumlah instances: 32427
- Jumlah atribut: 32
- Correctly Classified Instances: 30407 (93.7706 %)
- Incorrectly Classified Instances: 2020 (6.2294 %)

Naïve Bayes (Data Train)
- Tahap pengujian dilakukan dengan menggunakan 10 fold cross validation.
- Jumlah instances: 32427
- Jumlah atribut: 32
- Correctly Classified Instances: 25999 (80.1745 %)
- Incorrectly Classified Instances: 6429 (19.8255 %)

Naïve Bayes (Data Uji)
- Tahap pengujian dilakukan dengan menggunakan 10 fold cross validation.
- Jumlah instances: 32427
- Jumlah atribut: 32
- Correctly Classified Instances: 31235 (96.3241 %)
- Incorrectly Classified Instances: 1192 (3.6759 %)

J48 (Data Train)
- Tahap pengujian dilakukan dengan menggunakan 10 fold cross validation.
- Jumlah instances: 32427
- Jumlah atribut: 32
- Correctly Classified Instances: 26541 (81.8459 %)
- Incorrectly Classified Instances: 5887 (18.1541 %)

J48 (Data Uji)
- Tahap pengujian dilakukan dengan menggunakan 10 fold cross validation.
- Jumlah instances: 32427
- Jumlah atribut: 32
- Correctly Classified Instances: 32127 (99.0748 %)
- Incorrectly Classified Instances: 300 (0.9252 %).

TABEL 2

HASIL PENGUJIAN

	Naïve Bayes	J48	Random Forest Tree
Data (fold cross validation)	10		
Jumlah instances	32428		
Jumlah atribut	32		
Tingka akurasi	80,1745%	81,8459%	99,3802%
Prediksi Data	31235	32127	30407

Tebak untuk customer			

Berdasarkan hasil yang diperoleh pada tabel 2, didapatkan bahwa algoritma J48 memiliki akurasi yang lebih tinggi dibandingkan dengan Naïve Bayes. Percobaan dilakukan dengan menggunakan pengaturan testing yang Sama (10 fold cross validation) dengan jumlah atribut data dan instances yang sama. Akurasi yang diperoleh dari Naïve Bayes adalah sebesar 80.1745%, sedangkan dengan J48 didapatkan akurasi sebesar 81, 8459%.

Dengan menggunakan Naïve Bayes, hasil prediksi data tebak_class untuk customer yang Akan diberikan voucher adalah sebanyak 31235 customer. Sedangan dengan J48, prediksi untuk customer yang Akan diberikan voucher adalah sebanyak 32127 customer.
Berdasarkan hasil percobaan, disimpulkan bahwa dalam kasus dataset ini, Algoritma J48 lebih akurat daripada Naïve Bayes.
Namun, jika dibandingkan dengan algoritma Random Forest Tree, didapatkan bahwa algoritma Random Forest Tree memiliki akurasi yang lebih tinggi dibandingkan dengan Naïve Bayes. Percobaan dilakukan dengan menggunakan pengaturan testing yang Sama (10 fold cross validation) dengan jumlah atribut data dan instances yang sama. Akurasi yang diperoleh dari Naïve Bayes adalah sebesar 80.1745%, sedangkan dengan Random Forest didapatkan akurasi sebesar 99.3802 %.
Dengan menggunakan Naïve Bayes, hasil prediksi data tebak_class untuk customer yang Akan diberikan voucher adalah sebanyak 31235 customer. Sedangan dengan Random Forest, prediksi untuk customer yang Akan diberikan voucher adalah sebanyak 30407 customer.

Berdasarkan hasil percobaan, disimpulkan bahwa dalam kasus dataset ini, Algoritma Random Forest lebih akurat daripada Naïve Bayes.
Berdasarkan hasil yang diperoleh, didapatkan bahwa algoritma Random Forest memiliki akurasi yang lebih tinggi dibandingkan dengan J48. Percobaan dilakukan dengan menggunakan pengaturan testing yang Sama (10 fold cross validation) dengan jumlah atribut data dan instances yang sama. Akurasi yang diperoleh dari J48 adalah sebesar 81.8459 %, sedangkan dengan Random Forest didapatkan akurasi sebesar 99.3802 %.
Dengan menggunakan Naïve Bayes, hasil prediksi data tebak_class untuk customer yang Akan diberikan voucher adalah sebanyak 32127 customer. Sedangan dengan Random Forest, prediksi untuk customer yang Akan diberikan voucher adalah sebanyak 30407 customer.
Berdasarkan hasil percobaan, disimpulkan bahwa dalam kasus dataset ini, Algoritma Random Forest lebih akurat daripada Naïve Bayes.

5. Kesimpulan

Berdasarkan analisis penggunaan data mining baik terhadap data proses maupun kualitas, dapat ditarik kesimpulan sebagai berikut:

1. Dalam kasus dataset ini, algoritma Random Forest Tree lebih akurat daripada Naïve Bayes, dan J48
2. Dengan software Weka, dari 32427 *records* yang diolah, sebanyak 30407 data (99.3802%) dinyatakan benar sehingga hasil pengetahuan dalam bentuk pohon keputusan (*decision tree*) dapat digunakan untuk membantu memberikan alternatif rekomendasi untuk melakukan peningkatan loyalitas pelanggan dengan intelligent voucher.
3. Preprocessing, Filtering, dan Normalisasi dataset penting dalam konsep data mining. Preprocessing dan Filtering yang baik dapat meningkatkan akurasi prediksi. Sedangkan normalisasi digunakan untuk pengelompokan data yang memiliki range besar sehingga jangkauannya dapat diperkecil.

Daftar Pustaka

[1] S. D. H. Permana, "E-marketing strategy in game industry with social media using e-business model," pp. 258–263, 2016.

[2] B. C. Hartanto *et al.*, "Perancangan dan Pembuatan Website E-Commerce untuk UMKM yang dibina oleh Universitas Kristen Petra," pp. 1–6, 2017.

[3] M. A. Ghofar and Y. I. Kurniawan, "Aplikasi Pengelompokan Pelanggan Pada Ums Store Menggunakan Algoritma K-Means," *J. Teknol. Manaj. Inform.* –, vol. 4, no. 1, 2018.

[4] H. Hong *et al.*, "Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China)," *Catena*, vol. 163, no. January, pp. 399–413, 2018.

[5] G. Gunadi and D. I. Sensuse, "Penerapan Metode Data Mining Market Basket Analysis Terhadap Data Penjualan Produk Buku Dengan Menggunakan Algoritma Apriori Dan Frequent Pattern Growth (Fp-Growth) :," *Telematika*, vol. 4, no. 1, pp. 118–132, 2012.

[6] G. Gerber *et al.*, “The impact of churn on client value in health insurance , evaluation using a random forest under random censoring". HAL Id : hal-01807623, 2018.

[7] A. Fikri, “Penerapan Data Mining Untuk Mengetahui Tingkat Kekuatan Beton Yang Dihasilkan Dengan Metode Estimasi Menggunakan Linear Regression,” *Fak. Ilmu Komput. UDINUS*, pp. 1–12, 2013.

[8] Mediana Aryuni, “TAHAP-TAHAP DATA MINING,” *Bina Nusantara*, 2016. .