Uji Algoritma Random Forest Pada Dataset Online Shoppers Purchasing Intention

Arif Purnama[1], Ali Maulana Yusup[2], Agung Wibowo[3], Desi Susilawati[4]

[1,2] Universitas Bina Sarana Informatika
Jl. Cemerlang No. 8 Sukakarya, Sukabumi
E-mail : arifpurnama6@gmail.com[1], arimaulanayusup071198@gmail.com[2],
agung.awo@bsi.ac.id[3], desi.dlu@bsi.ac.id[4]

ABSTRAK

Online shopper merupakan pembelanjaan online yang saat ini sedang maraknya dilakukan oleh hampir semua orang. Belanja secara *online* dapat mempermudah proses transaksi tanpa harus bertemu langsung dengan penjual. Banyak penelitian yang sudah dilakukan dalam menguji *online shopper purchasing intention* salahsatunya penelitian yang dilakukan oleh C. Okan Sakar pada tahun 2018 yang menggunakan metode naive bayes dan *random forest.* beberapa penelitian sebelum nya dapat disimpulkan bahwa dataset yang digunakan sebagian besar tidak dilakukan Resample terlebih dahulu. Untuk memperoleh akurasi yang lebih tinggi dengan Random Forest maka dilakukan *Resample* dataset menjadi 2 bagian yakni untuk *Training* sebesar 60% dari dataset dan 40% untuk *Testing*. Hasilnya dengan melakukan Resample dataset terlebih dahulu maka dapat disimpulkan membuat akurasi menjadi lebih tinggi daripada dataset yang tidak dilakukan Resample terlebih dahulu. Dengan dilakukan komparasi juga menunjukan bahwa Random Forest memperoleh hasil yang lebih baik dibandingan algoritma lain.

Kata kunci : ***Online Shopper, Online Shopper Purchasing*** **Intention, Algoritma, Data Mining, Dataset,** ***Random Forest***

ABSTRACT

Online shopper is an online shopping that is currently being done by almost everyone. Shopping online can simplify the transaction process without having to meet directly with the seller. Many studies have been conducted in testing online shopper purchasing intention, one of which is research conducted by C. Okan Sakar in 2018 which used the naive bayes and random forest methods. Several previous studies concluded that most of the datasets used were not resample first. To obtain higher accuracy with Random Forest, Resample the dataset into 2 parts, namely for Training 60% of the dataset and 40% for Testing. As a result, by resample the dataset first, it can be concluded that the accuracy is higher than the dataset that was not resample first. By doing the comparisons, it also shows that Random Forest gets better results than other algorithms.

Keyword : Online Shopper, Online Shopper Purchasing Intention, Algorithm, Data Mining, Dataset, Random Forest

1. PENDAHULUAN

Dataset *Online Shoppers Purchasing Intention* merupakan penelitian yang dirancang untuk mengukur pengguna dalam menyelesaikan transaksi yang bertujuan untuk menawarkan konten bagi meraka yang berniat untuk membeli. Dataset OSPI terdiri dari 12.330 data dan memiliki 18 atribut yang terbagi menjadi dua bagian yakni Fitur Numerikal yang digunakan dalam model analisis pelaku pengguna sedang Fitur Kategorikal digunakan dalam model analisis perilaku pengguna. (C. Okan Sakar S. O., 2018).

Tabel 1. Fitur Numerikal

Atribute name	*Atribute description*	*Min value*	*Max value*	*Sd*
Administrative	*Number of pages visited by the visitor about account management*	*0*	*27*	*3.32*
Administrative duration	*Total amount of time (in seconds) spent by the visitor on account management related pages*	*0*	*3398*	*176.70*
Informational	*Number of pages visited by the visitor about web site, communication and address information of the shopping site*	*0*	*24*	*1.26*
Informational duration	*Total amount of time (in seconds) spent by the visitor on informational pages*	*0*	*2549*	*140.64*
Product related	*Number of pages visited by visitor about product related pages*	*0*	*705*	*44.45*
Product related duration	*Total amount of time (in seconds) spent by the visitor on product related pages*	*0*	*63.973*	*1912.25*
Bounce rate	*Average bounce rate value of the pages visited by the visitor*	*0*	*0.2*	*0.04*
Exit rate	*Average exit rate value of the pages visited by the visitor*	*0*	*0.2*	*0.05*
Page value	*Average page value of the pages visited by the visitor*	*0*	*361*	*18.55*
Special day	*Closeness of the site visiting time to a special day*	*0*	*1.0*	*0.19*

Tabel 2. Fitur Kategorikal

Atribute name	*Atribute description*	*Number of categorical values*
Operatiing system	*Operating system of the visitor*	*8*
Browser	*Browser of the visitor*	*13*
Region	*Geographic region from which the session has been started by the visitor*	*9*
Traffic type	*Traffic source by which the visitor has arrived at the web site (e.g., banner, sms, direct)*	*20*
Visitortype	*Visitor type as ‘‘new visitor,’’ ‘‘returning visitor,’’ and ‘‘other’*	*3*
Weekend	*Boolean value indicating whether the date of the visit is weekend*	*2*
Month	*Month value of the visit date*	*12*
Revenue	*Class label indicating whether the visit has been finalized with a transaction*	*2*

Random Forest merupakan salah satu metode yang digunakan untuk klasifikasi dengan membangun banyak pohon klasifikasi. RF dapat meningkatkan akurasi karena adanya pemilihan secara acak dalam membangkitkan simpul anak untuk setiap node dan diakumulasi hasil klasifikasi dari setiap pohon kemudian dipilih klasifikasi yang paling banyak muncul (Zainal, 2016).

Data sumber penelitian yang digunakan dalam pengujian akurasi *Online Shoppers Purchasing Intention* menggunakan data publik dari *UCI Machine Learning Repository*.

Tabel 3. Jurnal yang telah melakukan penelitian tentang dataset OSPI

Judul	Penulis	Tahun	Metode	Akurasi
An Intelligent Apparel Recommendation System For Online Shopping Using Style Classification	C. Perkinian and P. Vikkraman	2015		
Intelligent Decision Support for Data Purchase	Denis Mayr Lima Martins, Gottfried Vossen, Fernando Buarque de Lima Neto	2017		
Real-time prediction of online shoppers’ purchasing intention	C. Okan Sakar	2018	Naive Bayes	88.92 %

using multilayer perceptron and LSTM recurrent neural networks			Random Forest	89.51 %
Comparison of Machine Learning Algorithms Using WEKA and Sci-Kit Learn in Classifying Online Shopper Intention	Yefta Christian	2019	J48	89.294%
			Naive Bayes	80.886%
			MLP	88.565%
			SVM	88.078%
			Random Forest	90.133%
Prediction Commercial Intent of Online Consumers using Machine Learning Techniques	Mete Alpaslan Katircioglu	2018	C45	88.92 %
			Random Forest	89.51 %
			SVM	88.25 %
			Multilayer	87.45 %

C. Okan Sakar (2018) menguji Dataset OSPI menggunakan algoritma *Naive Bayes, Random Forest* pada *Dataset Online Shoppers Purchasing Intention* dan menunjukan bahwa *Random Forest* memperoleh akurasi tertinggi sebesar 89.51% dibandingkan algoritma yang lain. Hasil uji akurasi *Random Forest* memiliki akurasi tertinggi didukung oleh hasil dari pengujian yang di publikasikan oleh Yefta Christian (2019) dan Mete Alpaslan Katircioglo (2018) menyatakan bahwa akurasi tertinggi adalah algoritma *Random Forest*.

2. METODOLOGI

Metode Random Forest adalah pengembangan dari metode CART, yaitu dengan menerapkan metode bootstrap aggregating (bagging) dan random feature selection. Dalam random forest, banyak pohon ditumbuhkan sehingga terbentuk hutan (forest), kemudian analisis dilakukan pada kumpulan pohon tersebut (Adnyna, 2015).. Pada gugus data yang terdiri atas n amatan dan p peubah penjelas, random forest dilakukan dengan cara :

a) Lakukan penarikan contoh acak berukuran n dengan pemulihan pada gugus data. Tahapan ini merupakan tahapan bootstrap.
b) Dengan menggunakan contoh bootstrap, pohon dibangun sampai mencapai ukuran maksimum (tanpa pemangkasan). Pada setiap simpul, pemilihan pemilah dilakukan dengan memilih m

peubah penjelas secara acak, dimana m << p. Pemilah terbaik dipilih dari m peubah penjelas tersebut. Tahapan ini adalah tahapan random feature selection.

c) Ulangi langkah 1 dan 2 sebanyak k kali, sehingga terbentuk sebuah hutan yang terdiri atas k pohon.

3. LANDASAN TEORI

Data Mining

Data mining merupakan bagian dari tahapan proses Knowledge Discovery in Database (KDD). Dengan data mining, kita dapat melakukan pengklasifikasian, memprediksi, memperkirakan dan mendapatkan informasi lain yang bermanfaat dari kumpulan data dalam jumlah yang besar (Mardi, 2018).

Online Shopper Intention

online shopper intention bertujuan untuk memprediksi apakah pengguna menghasilkan pendapatan atau tidak. Salah satu cara untuk mendapatkan informasi atau pola dari kumpulan data yang besar adalah dengan menggunakan teknik-teknik dalam data mining (Ardiyansyah et al., 2018).

Random Forest

Klasifikasi algoritma random forest adalah teknik pembelajaran machine learning yang menghasilkan klasifikasi dalam bentuk forest (hutan) dari decision trees. Random forest memiliki banyak tree dan setiap tree ditanam dengan cara yang sama. Tree dengan variabel x akan ditanam sejauh mungkin dengan tree dengan variabel y. Dalam perkembangannya, sejalan dengan bertambahnya dataset, maka tree pun akan ikut berkembang. (M. Salim, 2016).

4. HASIL DAN PEMBAHASAN

Pengujian yang dilakukan ini menggunakan algoritma *Random Forest* untuk menghasilkan akurasi yang lebih baik lagi. Data yang digunakan dalam pengujian ini menggunakan *Dataset Online Shoppers Purchasing Intention* dari UCI Repository dengan link sebagai berikut https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset. Pada dataset *Online Shoppers Purchasing Intention* memiliki data sebanyak 12.330. dengan rincian 125 data memiliki data yang sama atau *Duplicate.* Maka untuk membuat tidak ada data yang sama dalam dataset dilakukanlah *Remove Duplicate* sehingga menghasilkan 12.205 data.

Data diuji menggunakan aplikasi *Waikato Enviroment For Knowledge Analysis (WEKA)* Metode pengujian ini menggunakan *10-Fold-Cross-Validation* dengan membagi data menjadi 10 bagian dimana bagian pertama digunakan untuk pelatihan dan sembilan bagian lainnya digunakan untuk pengujian. Pengujian ini menggukan Laptop dengan Prosesor i5-7200U 2.50Ghz dan Memori 8GB.

Dalam dataset ini tidak terdapat *Missing Value* maka Setelah dilakukan *Remove Duplicate* maka langsung dilakukan *Resample* dataset dengan membagi dataset menjadi dua bagian yakni *Training* dan *Testing*

dengan perhitungan sebesar 60% untuk *Training* dan 40% untuk *Testing*, sehingga mendapatkan hasil sebanyak 7323 atau 60% yang digunakan nanti untuk *Training* dan sebanyak 4882 atau 40% akan digunakan untuk *Testing*. Setelah dillakukan training terhadap dataset training dengan menggunakan *10 Folds-Cross-Validation* maka didapat hasil sekitar 94.95%, untuk melakukan perbandingan bahwa Random Forest merupakan Algoritma yang memiliki tingkat akurasi tertinggi, kami membandingkannya dengan Algoritma *Naive Bayes* dengan rincian sebagai berikut:

Tabel 4. Perbandingan Random Forest dan Naive Bayes

Metode	Akurasi	F1-Score	Kappa Statistic	MAE
Random Forest	94.95%	0.949	0.8016	0.0986
Naive Bayes	84.95%	0.845	0.4155	0.1655

Untuk melakukan training terhadap dataset Training (60%) maka dilakukan dilakukan *Re-Evaluate Model On Current Test Set* maka didapat hasil akurasi sekitar 93.87% dengan rincian sebagai berikut:

Tabel 5. *Re-Evaluate Model On Current Test Set*

Akurasi	F1-Score	Kappa Statistic	MAE
93.87%	0.937	0.765	0.1045

Komparasi

Pada proses ini dilakukan komparasi pada Percent Correct Dan Mean Absolute Error dengan hasil sebagai berikut :

a. Percent Correct

```
Dataset                 (1) trees.Ra | (2) trees
(3) bayes
----------------------------------------------
---------
online_shoppers_intention(100)
94.89 |   91.90 *   81.01 *
online_shoppers_intention(100)
90.15 |   89.17 *   81.88 *
----------------------------------------------
---------                          (v/ /*) |
(0/0/2)   (0/0/2)
```

Perbandingan ketiga algoritma menunjukan bahwa Random Forest mendapat nilai yang lebih besar dibandingkan yang lain

b. Mean Absolute Error

```
Dataset                 (1) trees.R | (2) tree
(3) bayes
----------------------------------------------
---------
online_shoppers_intention(100)   0.10
|   0.11 v   0.24 v
online_shoppers_intention(100)   0.14
|   0.14     v   0.22 v
----------------------------------------------
---
(v/ /*) |  (1/1/0)  (2/0/0)
```

5. KESIMPULAN

Dari hasil testing yang kami lakukan, beberapa penelitian sebelum nya dapat disimpulkan bahwa dataset yang digunakan sebagian besar tidak dilakukan Resample terlebih dahulu. Untuk memperoleh akurasi yang lebih tinggi dengan Random Forest maka dilakukan *Resample* dataset menjadi 2 bagian yakni untuk *Training* sebesar 60% dari dataset dan 40% untuk

Testing. Hasilnya dengan melakukan Resample dataset terlebih dahulu maka dapat disimpulkan membuat akurasi menjadi lebih tinggi daripada dataset yang tidak dilakukan Resample terlebih dahulu. Dengan dilakukan komparasi juga menunjukan bahwa Random Forest memperoleh hasil yang lebih baik dibandingan algoritma lain.

DAFTAR PUSTAKA

Adnyana, I Made Budi. (2015). Prediksi Lama Studi Mahasiswa Dengan Metode Random Forest (Studi Kasus : Stikom Bali). CSRID Journal, Vol.8 No.3 Oktober 2015.

Ardiyansyah, P. A. Rahayuningsih, and Reza Maulana. (2018). Analisis Perbandingan Algoritma Klasifikasi Data Mining Untuk Dataset Blogger Dengan Rapid Miner. J. Khatulistiwa Inform., vol. VI, no. 6, pp. 20–28

Azizah, N. (2017). *Implementasi Dan Analisa Waktu Komputasi Pada Algoritma Random Forest Dengan Parallel Computing Di R.* Depok: repository.upi.edu.

Breiman, L. (2001). *Random Forest.* Kluwer Academic Publisher.

C. Okan Sakar, S. O. (2018). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *The Natural Computing Applications Forum 2018*.

C. Okan Sakar, Y. K. (2018). *Online Shoppers Purchasing Intention Dataset Data Set*. Retrieved from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

C. Perkinian, P. V. (2015). An Intelligent Apparel Recommendation System For Online Shopping Using Atyle Classification. *I J A B E R, Vol 13, No. 2*, 671-686.

Comparison of Machine Learning Algorithms Using WEKA and Sci-Kit. (2019). *JITE (Journal of Informatics and Telecomunication Engineering)*, 58-66.

Dennis Mayr Lima Martins, G. V. (2017, Agust 23-26). Intelligent Decision Support for Dataset. *WI*.

Ho, T. K. (1995). Random Decision Forest. *Montreal* (pp. 14-16). QC: Proceedings of the 3rd International Conference on Document Analysis and Recognition.

Katircioglu, M. A. (2018). Prediction Commercial Intent of Online Consumers using Machine Learning Techniques.

Lima, Y. J. (2016). *7th International Economics & Business Management Conference* (pp. 401-410). Procedia Economics and Finance.

Mardi, Yuli. (2018). Data Mining : Klasifikasi Menggunakan Algoritma C4.5. Jurnal Edik Informatika

Salim, M., (2016), Klasifikasi Tutupan Lahan Perkotaan Menggunakan Naïve Bayes Berbasis Forward Selection, Jurnal Teknosains, 10 (2): 165-182.

Zainal, A. M. (n.d.). (2016) *Ensemble of One Class Classifier for Network Ensemble of One Class Classifier for Network Information System.* Universiti Teknologi Malaysia.