SNESTIK

Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika

https://ejurnal.itats.ac.id/snestik dan https://snestik.itats.ac.id

Informasi Pelaksanaan :
SNESTIK III - Surabaya, 11 Maret 2023
Ruang Seminar Gedung A, Kampus Institut Teknologi Adhi Tama Surabaya

Informasi Artikel:
DOI :10.31284/p.snestik.2023.4040

Prosiding ISSN 2775-5126

Fakultas Teknik Elektro dan Teknologi Informasi-Institut Teknologi Adhi Tama Surabaya
Gedung A-ITATS, Jl. Arief Rachman Hakim 100 Surabaya 60117 Telp. (031) 5945043
Email : snestik@itats.ac.id

Perbandingan Model Logistic Regression dan *K*-Nearest Neighbors Dalam Prediksi Pembatalan Hotel

Mohammad Fahry Sholahuddin, Abdul Holik , Chelvin Suprapto, Iqbal Izha Mahendra, Sadewa Wibawanto, Muchamad Kurniawan

Institut Teknologi Adhi Tama Surabaya
e-mail: fahrysholahuddin@gmail.com

ABSTRACT

Predicting hotel cancellations is an important part of today's hotel revenue management system. With accurate prediction, it can be used as a reference to improve hotel performance. This research focuses on the comparison between two supervised learning methods, namely k-nearest neighbors and Logistic Regression, to predict hotel cancellations. We used K-Fold Cross Validation to evaluate both methods and found that k-nearest neighbors has higher accuracy, precision and recall values compared to Logistic Regression, with an accuracy value of 0.81, precision of 0.81 and recall of 0.80. The results of this study can help hotels improve revenue management and reduce the impact of room cancellations.

Keywords: *Hotel cancellations; k-nearest neighbors; Logistic Regression; Revenue management; Supervised learning methods*

ABSTRAK

Prediksi pembatalan hotel memiliki arti penting dalam sistem manajemen pendapatan hotel saat ini. Dengan prediksi yang akurat, maka dapat digunakan sebagai acuan dalam meningkatkan kinerja hotel. Penelitian ini berfokus pada perbandingan antara dua metode *supervised learning*, yaitu k-nearest neighbors dan Regresi Logistik, untuk memprediksi pembatalan hotel. Kami menggunakan K-Fold Cross Validation untuk mengevaluasi kedua metode tersebut dan menemukan bahwa *k*-nearest neighbors memiliki nilai akurasi, presisi, dan *recall* yang lebih tinggi dibandingkan dengan Regresi Logistik, dengan nilai akurasi

sebesar 0.81, presisi 0.81, dan *recall* 0.80. Hasil penelitian ini dapat membantu hotel dalam meningkatkan manajemen pendapatan dan mengurangi dampak pembatalan kamar.

Kata kunci: Manajemen pendapatan; Metode *supervised learning*; *k*-nearest neighbors; Prediksi pembatalan hotel; Regresi Logistik.

PENDAHULUAN

Prediksi pembatalan hotel memiliki pentingnya dalam sistem manajemen pendapatan hotel saat ini. Dengan prediksi yang akurat, dapat digunakan sebagai acuan dalam meningkatkan kinerja hotel. Penelitian dilakukan untuk menyelesaikan masalah ini dengan menggunakan berbagai pendekatan pembelajaran mesin.

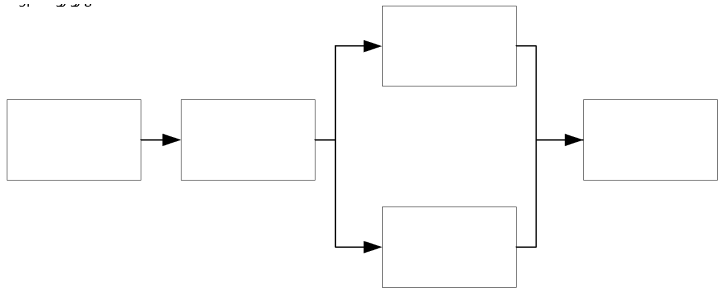
Pada penelitian sebelumnya oleh Kurniawan & Barokah, 2020, [3] digunakan metode *k*-NN dalam klasifikasi pengajuan kartu kredit. Data yang digunakan adalah data pengajuan kartu kredit. Hasil evaluasi metode *k*-NN pada penelitian tersebut menghasilkan nilai *precision* sebesar 92%, nilai *recall* sebesar 83%, dan nilai *accuracy* sebesar 93%.

Sedangkan dalam penelitian Putra & Azhar, 2021 [4] dengan judul Perbandingan Model Logistic Regression dan *Artificial Neural Network* pada Prediksi Pembatalan Hotel. Dalam penelitian ini membandingkan beberapa metode *supervised learning* diantaranya adalah Logistic Regression dan ANN. Dari penelitian yang dilakukan, metode ANN menghasilkan nilai *accuracy* sebesar 79.24%, *precision* sebesar 85.86% dan recall sebesar 53%. Sedangkan pada metode (Logistic Regression with GridSearchCV) menunjukan hasil yang lebih optimal dalam melakukan prediksi pembatalan pemesanan hotel, dengan nilai *accuracy* sebesar 79.77%, nilai *precision* sebesar 85.86% dan nilai *recall* sebesar 55.07%.

Dari hasil penelitian sebelumnya yang telah dilakukan, peneliti tertarik untuk melakukan penelitian untuk membandingkan beberapa metode *supervised learning* antara metode *k*-NN dan Logistic Regression. Hal tersebut bertujuan untuk mengetahui dari kedua metode tersebut mana yang memiliki nilai *accuracy*, *precision*, *recall* terbesar dalam memprediksi pembatalan hotel.

METODE

Alur proses yang akan dilakukan pada penelitian ini terdiri dari *Collect Dataset*, *Preprocessing, Modelling with k-nearest neighbors, Modeling with Logistic Regression, Evaluation.* Seperti yang terlihat pada Gambar 1.



Gambar 1. Diagram Alur Penelitian

Collect Dataset

Dataset dalam penelitian kali ini menggunakan *dataset* yang bersumber dari penelitian Antonio., 2019 [5] yang terdiri 119390 baris dengan 32 *independent* variables, 1 *dependent* variable. Seperti pada Tabel 1.

Tabel 1. Nama Kolom dan Tipe Data dari Independent Variables

Nama kolom	Tipe Data
hotel	object
lead_time	Numeric
arrival_date_year	Numeric
arrival_date_month	object
arrival_date_week_number	Numeric
arrival_date_day_of_month	Numeric
stays_in_weekend_nights	Numeric
stays_in_week_nights	Numeric
adults	Numeric
children	float64
babies	Numeric
meal	object
country	object
market_segment	object
distribution_channel	object
is_repeated_guest	Numeric
previous_cancellations	Numeric
previous_bookings_not_canceled	Numeric
reserved_room_type	object
assigned_room_type	object
booking_changes	Numeric
deposit_type	object
agent	float64
company	float64
days_in_waiting_list	Numeric
customer_type	object
adr	float64
required_car_parking_spaces	Numeric
total_of_special_requests	Numeric
reservation_status	object
reservation_status_date	object

Untuk nama kolom dan tipe data dari *dependent variable* adalah seperti pada Tabel 2.

Tabel 2. Nama Kolom dan Tipe Data dari Dependant Variable

Nama Kolom	Tipe Variabel
is_canceled	Numeric

Pre-Processing

Dataset yang telah diperoleh akan diolah terlebih dahulu melalui proses *pre-processing*. Proses ini meliputi penghapusan data duplikat, penggantian nilai *null* dengan nol, penghapusan kolom yang tidak berguna, dan normalisasi variabel numerik.

***K*-Nearest Neighbors**

k-nearest neighbors (k-NN) adalah metode pembelajaran *instance* dalam *supervised learning* yang termasuk dalam teknik *lazy learning*. Metode ini mencari *k* objek pada data latih yang memiliki kemiripan dengan objek pada data uji untuk melakukan klasifikasi. Algoritma *k*-NN mengklasifikasikan *instance* baru berdasarkan mayoritas jarak kedekatan dari kategori-kategori dalam *k*-NN. Sebuah sistem klasifikasi diperlukan untuk melakukan pencarian informasi menggunakan *k*-NN. [1]. Metode ini bekerja dengan berdasarkan pada jarak terpendek dari suatu sampel uji ke sampel latih dalam penentuan *k*-NN. Kemudian mayoritas dari *k*-NN tersebut diambil untuk dijadikan prediksi terhadap sampel uji. Jarak kedekatan atau jarak tetangga biasanya dihitung berdasarkan jarak Euclidean [1]. Langkah-langkah untuk menghitung metode *k*-nearest neighbors meliputi:

1. Menentukan parameter K
2. Menghitung jarak antara data training dan data testing.

   Perhitungan jarak yang paling umum dipakai pada perhitungan pada metode *k*-NN adalah menggunakan perhitungan jarak Euclidean seperti ditunjukkan oleh persamaan 1.

$$euc = \sqrt{\left(\sum_{i=1}^{n} (pi - qi)^2\right)} \quad (1)$$

   Keterangan :
   pi = data training
   qi = data testing
   i = variabel data
   n = dimensi data
3. Mengurutkan jarak yang terbentuk
4. Menentukan jarak terdekat sampai urutan *k*
5. Memasangkan kelas yang bersesuaian
6. Mencari jumlah kelas dari tetangga yang terdekat dan tetapkan kelas tersebut sebagai kelas data yang akan dievaluasi
7.

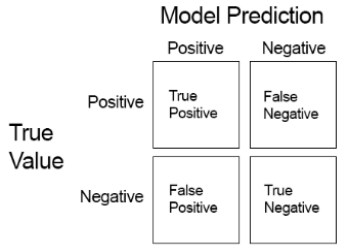
Logistic Regression

Logistic Regression adalah metode pengklasifikasian dalam statistical machine learning yang sering digunakan dalam data mining karena kinerjanya yang baik dalam memproses data berskala besar [2]. Logistic Regression menunjukkan keterkaitan antara *output* dalam bentuk pengklasifikasian biner terhadap variabel-variabel independen berdasarkan probabilitas dengan memprediksi nilai variabel dependen [6]. Adapun bentuk matematis untuk model regresi logistik ditunjukkan pada Persamaan 2, dimana $\sigma(t)$ merupakan fungsi logistik hasil adopsi dari aktivasi *sigmoid*, seperti yang dituliskan pada Persamaan 2.

$$\sigma(t) = \frac{1}{1+e^{-t}} \quad (2)$$

Evaluasi

Guna untuk melihat seberapa bagus model yang diajukan mampu memprediksi dan memperoleh nilai proporsi yang mampu diprediksi dengan tepat dibanding dengan nilai asli dari data yang tersedia, maka dapat dilihat dari nilai Confusion Matrix pada Gambar 2.



Gambar 2. matriks konfusi

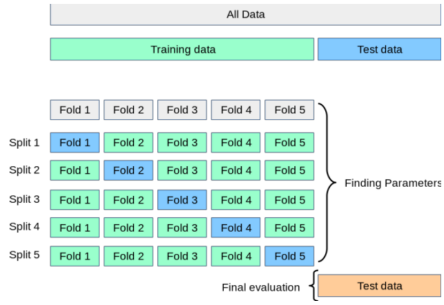
Gambar 2 menjelaskan konsep True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN) dalam evaluasi model. TP terjadi ketika keduanya sama-sama positif, TN ketika keduanya sama-sama negatif, FP ketika model positif tetapi data negatif, dan FN ketika model negatif tetapi data positif [7]. Berikut ini adalah rumus persamaan untuk mengukur nilai *accuracy* pada persamaan 3, precision pada persamaan 4, dan *recall* pada persamaan 5.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

Validasi model tidak hanya menggunakan Confusion Matrix, tetapi juga K-Fold Cross Validation. Metode ini membagi data ke dalam kelompok sebanyak nilai K-Fold dan satu kelompok digunakan sebagai data uji, sedangkan kelompok lainnya digunakan sebagai data latih Seperti yang ditunjukkan pada Gambar 3.



Gambar 3. Penerapan K-Fold Cross Validation

HASIL DAN PEMBAHASAN

Seperti yang telah dipaparkan sebelumnya bahwa tahapan yang dilakukan pada penelitian ini adalah dengan melakukan pembagian data training dan data testing, dengan pembagian K-Fold Cross Validation sebesar K-Fold=5, sehingga menjadi 80% data training dengan jumlah data 69784 dan 20% data testing dengan jumlah data 17446.

Dalam proses validasi pertama dilakukan K-Fold Cross Validation pada algoritma *k*-nearest neighbors, sehingga didapatkan hasil akhir berupa nilai akurasi dan hasil klasifikasi pada Tabel 3.

Tabel 3. Hasil K-Fold Cross Validation *k*-nearest neighbors

K-Fold Cross Validation	*Accuracy*	*Precision*	*Recall*
Pertama	0.81	0.80	0.79
Kedua	0.80	0.80	0.79
Ketiga	0.81	0.80	0.79
Keempat	0.80	0.80	0.80
Kelima	0.81	0.81	0.80

Dalam proses validasi kedua dilakukan K-Fold Cross Validation pada metode Logistic Regression, sehingga didapatkan hasil akhir berupa nilai akurasi dan hasil klasifikasi pada Tabel 4.

Tabel 4. Hasil K-Fold Cross Validation Logistic Regression

K-Fold Cross Validation	Accuracy	Precision	Recall
Pertama	0.80	0.79	0.80
Kedua	0.79	0.78	0.79
Ketiga	0.80	0.80	0.80
Keempat	0.78	0.77	0.79
Kelima	0.79	0.78	0.79

Dari proses K-Fold Cross Validation metode *k*-nearest neighbors didapat nilai Accuracy tertinggi dengan 0.81, Precision sebesar 0.81 Serta Recall sebesar 0.80. Sedangkan proses K-Fold Cross Validation metode Logistic Regression didapat nilai Accuracy tertinggi dengan 0.80, Precision sebesar 0.80 Serta Recall sebesar 0.80.

KESIMPULAN

Dari kedua hasil evaluasi melalui K-Fold Cross Validation yang telah dilakukan pada metode dan *k*-nearest neighbors dan Logistic Regression. Didapatkan hasil bahwa metode *k*-nearest neighbors mempunyai nilai *Accuracy*, *Precision* serta *Recall* tertinggi dalam proses prediksi pembatalan hotel dibandingkan dengan Logistic Regression. Dimana didapat nilai tertinggi dari proses K-Fold Cross Validation dengan *Accuracy* sebesar 0.81, *Precision* sebesar 0.81 serta *Recall* sebesar 0.80.

Peneliti menyimpulkan bahwa metode *k*-nearest neighbors memiliki hasil terbaik dalam prediksi pembatalan hotel dibandingkan dengan Logistic Regression. Peneliti merekomendasikan penambahan variabel dan perluasan sampel dalam penelitian selanjutnya untuk meningkatkan validitas hasil. Peneliti juga menyarankan penggunaan algoritma machine learning lainnya dan penggabungan kedua metode untuk meningkatkan akurasi prediksi.

DAFTAR PUSTAKA

[1] D. Cahyanti, A. Rahmayani, dan S. A. Husniar, "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," *Indones. J. Data Sci.*, vol. 1, no. 2, hal. 39–43, 2020, doi: 10.33096/ijodas.v1i2.13.

[2] Z. Yang, "Application of Logistic Regression with Filter in Data Classification," hal. 3755–3759, 2019, doi: 10.23919/ChiCC.2019.8865281.

[3] Y. I. Kurniawan dan T. I. Barokah, "Klasifikasi Penentuan Pengajuan Kartu Kredit Menggunakan k-nearest neighbors," *J. Ilm. Matrik*, vol. 22, no. 1, hal. 73–82, 2020, doi: 10.33557/jurnalmatrik.v22i1.843.

[4] M. S. T. Putra dan Y. Azhar, "Perbandingan Model Logistic Regression dan Artificial Neural Network pada Prediksi Pembatalan Hotel," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 6, no. 1, hal. 29–37, 2021, doi: 10.14421/jiska.2021.61-04.

[5] N. Antonio, A. de Almeida, dan L. Nunes, "Hotel booking demand datasets," *Data Br.*, vol. 22, hal. 41–49, 2019, doi: 10.1016/j.dib.2018.11.126.

[6] X. Zou, Y. Hu, Z. Tian, dan K. Shen, "Logistic Regression Model Optimization and Case Analysis," *Proc. IEEE 7th Int. Conf. Comput. Sci. Netw. Technol. ICCSNT 2019*, hal. 135–139, 2019, doi: 10.1109/ICCSNT47585.2019.8962457.

[7] R. Strandberg dan J. Låås, "A comparison between Neural networks, Lasso regularized Logistic regression, and Gradient boosted trees in modeling binary sales," 2019, [Daring]. Tersedia pada: https://www.diva-portal.org/smash/get/diva2:1319871/FULLTEXT02