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 II - Surabaya, 26 Maret 2022
Ruang Seminar Gedung A, Kampus Institut Teknologi Adhi Tama Surabaya

Informasi Artikel:
DOI : 10.31284/p.snestik.2022.2927

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

Seleksi Fitur pada Klasifikasi K-Nearest Neighbors untuk Data Churn for Bank Customers dengan Analisis Korelasi

Ika Maylani[1,*], Fadlur Rochman[2], Norma Devi Kurniasari[3]

Institut Teknologi Insan Cendekia Mandiri, Sidoarjo, Jawa Timur
e-mail: [1,*]*ika.maylani7@gmail.com*, [2]*im.fadlurrochman@gmail.com*, [3]*norma.devi@gmail.com*

ABSTRACT

One of the problems that arise in the data learning process is the large amount of data and many features involved. One technique that can be used to deal with this problem is feature selection with the aim of reducing the number of features. Approaches that can be used to perform feature selection include correlation analysis. Correlation analysis can be used to find out how influential a feature is on the results or classification targets. This study performs feature selection using correlation analysis and then tested by classifying the data using the K-Nearest Neighbor method. The data used is Churn for Bank Customers taken from Kaggle. The test results show that reducing the number of features based on a low correlation coefficient value can increase the accuracy. Features that are considered important are Age and IsActiveMember.

Keywords: *Churn for Bank Customers Data; Correlation analysis; Feature selection; K-Nearest Neighbor.*

ABSTRAK

Salah satu permasalahan yang muncul pada proses pembelajaran data yakni jumlah data yang besar dan banyaknya fitur yang dilibatkan. Salah satu teknik yang bisa digunakan untuk menangani hal tersebut yakni seleksi fitur dengan tujuan untuk mereduksi jumlah fitur. Pendekatan yang bisa digunakan dalam melakukan seleksi fitur antara lain analisis korelasi. Analisi korelasi dapat digunakan untuk mengetahui seberapa berpengaruh fitur terhadap hasil atau target klasifikasi. Penelitian ini melakukan seleksi fitur menggunakan analisis korelasi kemudian diuji dengan mengklasifikasikan data dengan memanfaatkan metode K-Nearest Neighbor. Data yang digunakan yakni data Churn for Bank Customers yang diambil dari Kaggle. Hasil uji coba menunjukkan bahwa pengurangan jumlah fitur berdasarkan nilai koefisien korelasi yang rendah dapat meningkatkan nilai akurasi. Fitur yang dianggap penting yakni Age dan IsActiveMember.

Kata kunci: Analisis Korelasi; Data Churn for Bank Customers; K-Nearest Neighbor; Seleksi fitur.

PENDAHULUAN

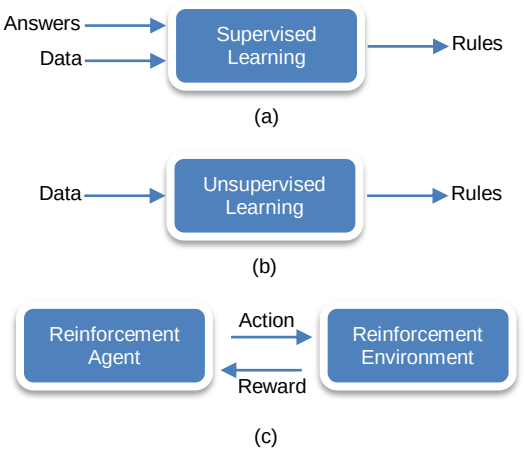
Secara umum, teknik pembelajaran mesin dapat dikategorikan menjadi terarah *(supervised)*, tak terarah *(unsupervised)*, dan *reinforcement learning* sebagaimana juga disebut dalam [1]. Teknik pembelajaran terarah biasanya akan memberikan hasil identifikasi yang bagus terutama saat data pembelajaran *(ground-truth)* yang disediakan mencukupi [2]. Oleh karena itu, baik-tidaknya data pembelajaran juga akan mempengaruhi hasil pembelajaran. Data pembelajaran ini umumnya mengandung data yang telah dilabeli dengan hasil identifikasi atau jawabannya. Ilustrasi masukan dan keluaran pada teknik-teknik pembelajaran mesin tersebut dapat dilihat pada Gambar 1.

Jenis pekerjaan yang bisa dilakukan dalam teknik pembelajaran terarah yakni klasifikasi dan regresi [2]. Dibandingkan dengan regresi, klasifikasi akan sangat cocok untuk diterapkan dalam membangun model yang berbasis pada data-data diskrit [3]. Salah satu metode klasifikasi yang biasa digunakan antara lain K-Nearest Neighbor (K-NN).

Metode K-NN ini cukup banyak diterapkan pada berbagai penelitian, di antaranya oleh Papernot dan McDaniel [4] yang menggunakan K-NN untuk mengembangkan metode Deep Learning; Saadatfar, dkk. [5] yang menggunakan K-NN untuk melakukan klasifikasi Big Data dengan mengombinasikan teknik *pruning* terhadap data; serta Ahuja, dkk. [6] yang menggunakan K-NN untuk membuat sistem rekomendasi *movie* dengan mengombinasikan bersama teknik *clustering*.

Selain membicarakan tentang metode klasifikasi, hal yang juga biasa dimunculkan oleh para peneliti yakni data yang digunakan. Beberapa penelitian menggunakan data primer yang diambil langsung dari sumber data, sebagian yang lain mengambil data yang sudah siap pakai. Salah satu penyedia data yang siap pakai tersebut yakni Kaggle.

Kaggle merupakan bagian dari Google yang menaungi komunitas praktisi pada bidang sains data dan pembelajaran mesin. Melalui Kaggle, para praktisi dapat mencari dan mempublikasikan data penelitian pada platform tersebut. Selain sebagai tempat berbagi data, Kaggle juga dapat dijadikan sebagai tempat untuk menemukan kompetisi-kompetisi atau tantangan-tantangan yang diselenggarakan, baik yang mengusung hadiah sebagai imbalannya maupun yang hanya mengusung pengembangan pengetahuan para pesertanya [7].



Gambar 1. (a) Teknik pembelajaran terarah, (b) Teknik pembelajaran tak terarah, dan (c) *Reinforcement learning*.

Sumber: Morocho-Cayamcela, dkk. [1]

Karakteristik sebuah data penelitian pada bidang mesin pembelajaran yakni jumlah baris data yang besar atau jumlah fitur yang banyak. Besarnya jumlah baris ataupun banyaknya jumlah fitur tersebut bisa mengakibatkan semakin lamanya komputasi atau cukup termakannya sumber daya komputer. Salah satu teknik yang bisa digunakan selain *pruning* pada data yakni seleksi fitur.

Terdapat beberapa teknik seleksi fitur, salah satunya menggunakan analisis korelasi, dalam hal ini analisi korelasi Pearson. Teknik ini pernah digunakan oleh Liu, dkk. [8] untuk menyeleksi fitur pada aktivitas harian berkaitan dengan sistem rumah cerdas. Selain itu, Rizqiwati, dkk. [9] juga pernah menggunakannya untuk menyeleksi fitur pada kasus analisis kelelahan berdasarkan rekaman data EEG responden. Penelitian lainnya, yakni Sugianela dan Ahmad [10] yang menggunakan teknik analisis korelasi untuk menyeleksi fitur pada sistem pendeteksi intrusi.

Penelitian ini mencoba untuk menerapkan analisis korelasi untuk mereduksi atau menentukan fitur penting pada data Churn for Bank Customers [11] yang diambil dari Kaggle. Fitur yang telah diseleksi akan diuji coba pada proses klasifikasi data menggunakan K-NN.

METODE

Tahapan yang dilakukan pada penelitian ini dapat dilihat pada Gambar 2. Data Churn for Bank Customers terdiri atas 14 fitur, termasuk target, yaitu: RowNumber, CustomerId, Surname, CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary, dan Exited dengan data sejumlah 10.000 baris. Data uji yang digunakan yakni sebesar 20%, sementara 80%-nya digunakan untuk data latih. Deskripsi detail untuk tiap fitur dapat dilihat pada Tabel 1.

Seleksi fitur awal sebagaimana pada Gambar 2 dimaksudkan untuk menanggalkan tiga fitur awal (RowNumber, CustomerId, dan Surname) yang tidak berdampak pada keluaran atau keputusan nasabah. *Encoding data kategorik* difungsikan untuk mengubah representasi data yang berjenis kategorik untuk memudahkan perhitungan saat pemrosesan lebih lanjut.

Gambar 2. Tahapan Penelitian

Tabel 1. Deskripsi Detail Fitur Data Churn for Bank Customers [11]

Fitur ke-	Nama Fitur	Deskripsi	Jenis Data
1	RowNumber	Nomor baris, tidak berpengaruh terhadap keluaran.	Kategorik
2	CustomerId	Berisi angka acak, tidak berpengaruh terhadap keluaran.	Kategorik
3	Surname	Nama belakang, tidak berpengaruh terhadap keluaran.	Kategorik
4	CreditScore	Skor kredit nasabah.	Numerik
5	Geography	Lokasi nasabah.	Kategorik
6	Gender	Jenis kelamin nasabah.	Kategorik
7	Age	Usia nasabah.	Numerik
8	Tenure	Lama menjadi nasabah.	Numerik
9	Balance	Saldo nasabah.	Numerik
10	NumOfProducts	Jumlah produk yang dibeli oleh nasabah dari bank.	Numerik
11	HasCrCard	Kepemilikan kartu kredit.	Kategorik
12	IsActiveMember	Keaktifan nasabah.	Kategorik
13	EstimatedSalary	Perkiraan gaji nasabah.	Numerik
14	Exited	Keputusan nasabah untuk meninggalkan bank atau tidak; Target klasifikasi.	Kategorik

Analisis korelasi digunakan untuk mendapatkan koefisien korelasi yang pada akhirnya difungsikan untuk menentukan fitur-fitur penting atas data. Tahapan terakhir yakni *Klasifikasi data dengan K-NN* dengan nilai K = 3. Tahapan ini bertujuan untuk membuktikan bahwa fitur-fitur yang kurang berkaitan dengan target layak untuk ditanggalkan dengan tetap mempertahankan performa klasifikasi. Jumlah fitur dijadikan sebagai salah satu bahan eksperimen untuk menentukan jumlah fitur yang representatif.

HASIL DAN PEMBAHASAN

Analisis korelasi dari fitur-fitur data terhadap target dapat dilihat dari koefisien korelasi yang dihasilkan. Hasil analisis tersebut dapat dilihat pada Gambar 3.

Gambar 3. Hasil Analisis Korelasi

Dapat dilihat pada Gambar 3, fitur setelah proses seleksi awal berjumlah sepuluh. Age merupakan fitur yang paling berkorelasi dengan tingkat pemutusan hubungan nasabah terhadap bank dengan nilai koefisien korelasi sebesar 0,285. Fitur dengan tingkat korelasi yang rendah, di bawah nilai tengah, yakni NumOfProducts; CreditScore; Tenure; EstimatedSalary; dan HasCrCard dengan masing-masing nilai koefisien korelasi 0,048; 0,027; 0,014; 0,012; dan 0,007.

Nilai tengah (median) dari daftar koefisien korelasi tersebut yakni 0,077. Nilai ini yang akan dijadikan sebagai pembatas *(threshold)* fitur-fitur yang akan digunakan dalam proses klasifikasi. Dengan kata lain, fitur yang akan digunakan pada proses klasifikasi yakni Age, IsActiveMember, Geography, Balance, dan Gender.

Hasil eksperimen proses klasifikasi K-NN dengan mengatur jumlah fitur dapat dilihat pada Tabel 2.

Tabel 2. Hasil Klasifikasi Data dengan K-NN

Eksperimen ke-	Jumlah Fitur	Fitur yang Dilibatkan	Akurasi
1	10	Age, IsActiveMember, Geography, Balance, Gender, NumOfProducts, CreditScore, Tenure, EstimatedSalary, dan HasCrCard	0,74
2	5	Age, IsActiveMember, Geography, Balance, dan Gender	0,73
3	4	Age, IsActiveMember, Geography, dan Balance	0,74
4	3	Age, IsActiveMember, dan Geography	0,80
5	2	Age dan IsActiveMember	0,81
6	1	Age	0,81

Berdasarkan data pada Tabel 2, eksperimen ke-1 melakukan klasifikasi dengan melibatkan seluruh fitur awal. Hal ini dimaksudkan untuk membandingkan dengan hasil pengurangan jumlah

fitur pada eksperimen berikutnya. Akurasi yang didapatkan oleh K-NN dengan melibatkan seluruh fitur yakni 0,74. Eksperimen ke-2 menghasilkan nilai akurasi 0,73, turun 0,01 poin dari eksperimen ke-1 yang melibatkan seluruh fitur. Akan tetapi, dapat dilihat pula bahwa dengan pengurangan fitur yang dilakukan secara bertahap, terjadi peningkatan akurasi. Nilai akurasi tertinggi dapat dicapai dengan jumlah fitur sebanyak 2 dan 1. Jumlah fitur yang hanya 1 memiliki nilai akurasi yang sama dengan nilai akurasi eksperimen ke-5 dengan jumlah fitur 2, dapat dikaitkan dengan nilai koefisien korelasi pada Gambar 3. Fitur Age memang memiliki korelasi yang sangat tinggi dibandingkan dengan fitur lainnya.

KESIMPULAN

Berdasarkan hasil uji coba, dapat disimpulkan bahwa seleksi fitur yang telah dilakukan berdampak pada peningkatan akurasi hasil klasifikasi. Adapun fitur yang paling mempengaruhi tingkat akurasi berdasarkan hasil uji coba dengan skenario yang telah disebutkan yakni Age dan IsActiveMember.

UCAPAN TERIMA KASIH

Peneliti mengucapkan terima kasih sebanyak-banyaknya kepada Institut Teknologi Insan Cendekia Mandiri, serta kepada Yayasan Yatim Mandiri Sidoarjo, Jawa Timur yang telah memberikan dukungan penuh kepada peneliti sehingga penelitian ini dapat terselesaikan dengan baik.

DAFTAR PUSTAKA

[1] M. E. Morocho-Cayamcela, H. Lee, dan W. Lim, "Machine learning for 5G/B5G mobile and wireless communications: Potential, limitations, and future directions," *IEEE Access*, vol. 7, hlm. 137184–137206, 2019, doi: 10.1109/ACCESS.2019.2942390.

[2] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, hlm. 44–53, Agu 2017, doi: 10.1093/nsr/nwx106.

[3] S. N. Shukla dan B. M. Marlin, "Interpolation-Prediction Networks for Irregularly Sampled Time Series," 2019.

[4] N. Papernot dan P. D. McDaniel, "Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning," *CoRR*, vol. abs/1803.04765, 2018, [Daring]. Tersedia pada: http://arxiv.org/abs/1803.04765

[5] H. Saadatfar, S. Khosravi, J. H. Joloudari, A. Mosavi, dan S. Shamshirband, "A New K-Nearest Neighbors Classifier for Big Data Based on Efficient Data Pruning," *Mathematics*, vol. 8, no. 2, 2020, doi: 10.3390/math8020286.

[6] R. Ahuja, A. Solanki, dan A. Nayyar, "Movie Recommender System Using K-Means Clustering AND K-Nearest Neighbor," dalam *2019 9th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, 2019, hlm. 263–268. doi: 10.1109/CONFLUENCE.2019.8776969.

[7] Kaggle, "Getting Started on Kaggle | Data Scxience Resources," 2022. https://www.kaggle.com/docs

[8] Y. Liu, Y. Mu, K. Chen, Y. Li, dan J. Guo, "Daily activity feature selection in smart homes based on pearson correlation coefficient," *Neural Processing Letters*, vol. 51, no. 2, hlm. 1771–1787, 2020.

[9] D. Risqiwati, A. D. Wibawa, E. S. Pane, W. R. Islamiyah, A. E. Tyas, dan M. H. Purnomo, "Feature Selection for EEG-Based Fatigue Analysis Using Pearson Correlation," dalam *2020 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 2020, hlm. 164–169. doi: 10.1109/ISITIA49792.2020.9163760.

[10] Y. Sugianela dan T. Ahmad, "Pearson Correlation Attribute Evaluation-based Feature Selection for Intrusion Detection System," dalam *2020 International Conference on Smart Technology and Applications (ICoSTA)*, 2020, hlm. 1–5. doi: 10.1109/ICoSTA48221.2020.1570613717.

[11] Kaggle, "Churn for Bank Customers." 2020. [Daring]. Tersedia pada: https://www.kaggle.com/mathchi/churn-for-bank-customers