e-ISBN 978-623-93343-3-8
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 20 April 2021

Pengaruh Seleksi Fitur Pada Algoritma *Machine Learning* Untuk Memprediksi Pembatalan Pesanan Hotel

I Gusti Naufhal Daffa Adnyana [1], Roihan Mufli Arjuna [2], Alfiyah Nur Indraini[3], Desta Sandya Pasvita [4]
S1 Informatika / Fakultas Ilmu Komputer
Program Studi Informatika, Universitas Pembangunan Nasional Veteran Jakarta
Jl. RS. Fatmawati Raya, Pd. Labu, Kec. Cilandak, Kota Depok, Jawa Barat 12450
igustinda@upnvj.ac.id , roihanma@upnvj.ac.id , alfiyahni@upnvj.ac.id , desta.sandya@upnvj.ac.id

Abstrak. Perkembangan media berbasis online mempengaruhi konsumen dalam melakukan pencarian informasi suatu produk dan jasa, salah satunya jasa pemesanan hotel dimana terdapat sistem *Booking*. Penelitian ini bertujuan untuk mengetahui pengaruh seleksi fitur terhadap performa algoritma machine learning dengan menggunakan dataset *Hotel Booking Demand* . Algoritma yang digunakan adalah *Decision Tree* dan *Random Forest* dengan menggunakan PCA untuk menemukan fitur yang berpengaruh dalam konsumen dalam melakukan pembatalan hotel. Hasil dari penelitian ini berupa nilai akurasi dari setiap algoritma dengan menggunakan seleksi fitur. Hasil akurasi dengan 3 , 5 dan 14 fitur menggunakan algoritma *Decision Tree* sebesar 0,954, 0,963, 0,971 dan *Random Forest* sebesar 0,963, 0,976, 0,980. Hasil penelitian tanpa menggunakan seleksi fitur dengan algoritma *Decision Tree* dan *Random Forest* sebesar 0.982 dan 0.987. Hasil analisis menunjukan dengan adanya seleksi fitur akan berpengaruh pada hasil dari akurasi setiap algoritma dimana pada penelitian ini justru menghasilkan akurasi yang lebih buruk dibandingkan tanpa seleksi fitur.

Kata Kunci: *random forest*, *decision tree*, PCA, seleksi fitur.

1 Pendahuluan

Perkembangan penggunaan internet yang pesat saat ini, menunjukkan adanya pergesaran teknologi yang semakin maju mengarah ke media berbasis online. Banyak konsumen yang memiliki kecenderungan untuk menelusuri (*surfing*) kelengkapan informasi produk dan jasa melalui internet dan melakukan pembelian secara online dikarenakan keterbatasan waktu serta kemudahan yang dirasakan. Salah satu nya dalam industri pariwisata dan travel yang sudah memanfaatkan teknologi dalam memesan tempat makan, tempat istirahat, transportasi, dan lain-lain. Dengan adanya teknologi yang membantu kegiatan dalam industri pariwisata dan travel ini selain memudahkan konsumen dalam memesan juga dapat membantu para pengusaha dalam mengelola pemasukan dengan baik dan benar sehingga dapat mengurangi kerugian yang mungkin terjadi dalam suatu perusahaan. Misalnya dalam pemesanan kamar hotel, dengan perkembangan teknologi terutama dalam bidang data science. Pengelola bisa mengetahui tingkah laku pelanggan (*customer*) dan memprediksi kemungkinan yang akan terjadi di masa mendatang. Misalnya dalam pemesanan kamar hotel, pengelola bisa mengetahui apakah pelanggan (*customer*) akan melakukan *cancel* atau pembatalan berdasarkan data-data yang ada. Dengan pihak pengelola mengetahui tingkah laku pelanggan (*customer*) dengan baik berdasarkan data maka pihak pengelola dapat memperbaiki layanannya sehingga dapat meminimalisir kerugian yang mungkin akan dialami. Selain meminimalisir kerugian, dengan mengetahui tingkah laku pelanggan (*customer*) maka pelanggan (*customer*) akan puas terhadap layanan yang diberikan sehingga para pelanggan (customer) ini akan merekomendasikan layanan yang telah diberikan kepada keluarga dan teman-temannya. Hal ini secara tidak langsung dapat meningkatkan pendapatan.

Pada penelitian ini algoritma yang digunakan untuk pemodelan pada data pemesanan hotel ini adalah algoritma *decision tree* dan *random forest* . Pada dasarnya kedua algoritma itu berupa *tree* / pohon. Algoritma yang paling simpel dalam *decision tree* yaitu ID3. ID3 merupakan algoritma dari *decision tree* yang membangun pohon keputusan secara rakus (*greedy*). Pohon keputusan yang dibuat oleh ID3 pertama dicari dahulu akar (*root*) dengan menghitung *information gain*, hasil perhitungan *information gain* yang terbesar akan menjadi akar (*root*).

e-ISBN 978-623-93343-3-8
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 20 April 2021

Sedangkan hasil perhitungan *information gain* yang lebih rendah akan menjadi cabang-cabang di bawahnya. Di akhir ID3 akan menghasilkan pohon keputusan yang mampu klasifikasi ataupun prediksi.

Algoritma *random forest* terdiri dari beberapa *decision tree* , semakin banyak pohon (*tree*) maka akan semakin akurat hasil klasifikasi yang dihasilkan dari *random forest* ini dan hasilnya tidak akan *overfitting*. Sebelum *random forest* membentuk pohon, beberapa sub dataset dibentuk dengan diambil data secara acak dengan sejumlah N sampel. N sampel tersebut harus lebih kecil dari jumlah N data asli. Proses itu berulang sampai mendapatkan *k* buah pohon. Kemudian solusi prediksi yang dihasilkan *random forest* ini dengan memilih kelas prediksi mayoritas.

Tujuan penelitian ini untuk mengetahui pengaruh seleksi fitur terhadap performa algoritma *machine learning* yaitu *decision tree* dan *random forest* dengan menggunakan teknik *Principal Component Analysis* (PCA) . PCA ini biasanya digunakan untuk mereduksi atau seleksi fitur pada data tanpa mengubah karakteristik data secara signifikan.

2 Landasan Teori

2.1 PCA (*Principal Component Analysis*)

PCA (Principal Component Analysis) adalah metode atau teknik yang paling sering digunakan untuk mereduksi atau menyederhanakan fitur pada data tanpa mengubah karakteristik dari data secara signifikan sehingga data masih dapat menggambarkan informasi pada data[1]. Teknik PCA akan mengekstraksi komponen utama kearah dimana data menggambarkan tingkat variabilitas yang maksimum [2]. PCA mentransfer data ke dalam sistem yang terkoordinasi oleh variabel yang tidak berkorelasi secara linier menggunakan teknik transformasi ortogonal[3].

2.2 *Ramdom Forest*

Random forest merupakan salah satu algoritma machine learning yang digunakan untuk kasus klasifikasi. Random forest dibuat dari beberapa *decision trees*, tiap *decision trees* akan tumbuh penuh serta tidak perlu proses pemotongan, semakin banyak *tree* / pohon semakin akurat hasilnya dan tidak akan *overfitting*[4].

Seperti yang telah disebutkan Random Forest terdiri dari beberapa *decision tree* atau bisa dikatakan Random Forest menghasilkan banyak pohon dari dataset. Sebelum Random Forest membentuk pohon, diambil data secara acak dari dataset dengan sejumlah N-sampel membentuk beberapa sub dataset atau istilahnya *bootstrap* . N-sampel tersebut harus lebih kecil dari jumlah N data asli. Proses tersebut terus berulang sampai mendapatkan *k* buah pohon secara acak. Kemudian proses pembelajaran model Random Forest akan menentukan solusi prediksi dengan cara memilih kelas prediksi mayoritas[5].

2.3 *Decision tree*

Algoritma yang paling simple dalam teknik decision tree adalah ID3. ID3 melakukan pencarian rakus sehingga tidak ada jaminan menghasilkan pohon keputusan yang optimum. Cara kerja ID3 dimulai dengan pohon kosong kemudian membangun pohon keputusan berdasarkan *information gain* dari atribut-atribut yang ada. Atribut yang memiliki *information gain* terbesar akan menjadi akar (*root*) pohon keputusan tersebut. Sedangkan atribut-atribut dengan *information gain* yang lebih rendah akan menjadi cabang-cabang di bawahnya. Di akhir ID3 akan menghasilkan pohon keputusan yang mampu mengklasifikasikan sampel-sampel data seakurat mungkin [5].

Tahap algoritma ID3 yaitu:

1. Pertama menghitung entropy dan *information gain* setiap atribut menggunakan rumus sebagai berikut

Rumus Entropy:

$$Entropy(S) = \sum_i^c p_i \ log_2 \ p_i \qquad \textbf{(1)}$$

Dimana c adalah jumlah kelas sedangkan p_i menyatakan porsi objek data (sampel) di kelas *i* terhadap jumlah semua sampel pada himpunan data

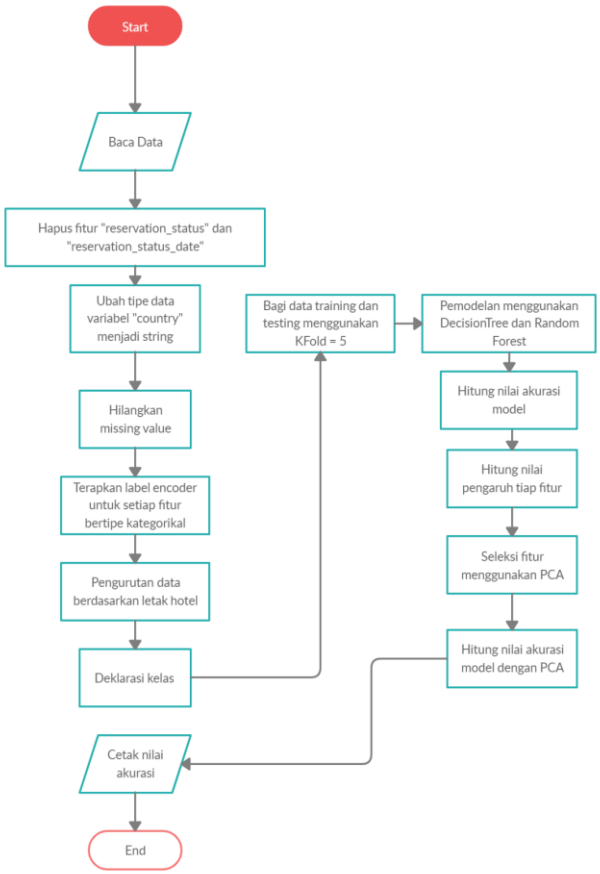
Rumus *Information Gain* :

$$Gain(S, A) = Entropy(S) - \sum_{v \epsilon Values(A)} \frac{|S_v|}{|S|} \ Entropy(S_v) \qquad \textbf{(2)}$$

Dimana A adalah atribut, V menyatakan suatu nilai yang mungkin untuk atribut A, *Values*(A) merupakan himpunan nilai-nilai yang mungkin untuk atribut A, $|S_v|$ adalah jumlah sampel untuk nilai v, $|S|$ menyatakan jumlah seluruh sampel data, dan Entropy(S) menyatakan entropy untuk sampel-sampel yang memiliki nilai v.

2. Menghitung information gain lagi sampai semua menghasilkan sebuah pohon keputusan. Atribut yang sudah dipilih tidak perlu lagi dimasukkan ke dalam perhitungan *information gain* [6].

3 Metodologi Penelitian



Gambar. 1. Alur Penelitian

1. Pengumpulan Data

Pada penelitian ini menggunakan dataset *Hotel Booking Deman* [7] yang berisi tentang data pemesanan hotel dari empat hotel yang berlokasi di Algarve, Portugal pada tahun 2015 yang dibagi dalam dua jenis yaitu *resort hotel* dan *city hotel*. Data tersebut digunakan pada penelitian sebelumnya oleh (Antonio, De Almeida, & Nunes, 2017) pada jurnal yang berjudul "*Predicting Hotel Booking Cancellation to Decrease Uncertainty and Increase Revenue*"[8]. Pada data yang digunakan dalam penelitian ini memiliki fitur sebanyak 32 fitur. Data akan disimpan dalam file hotel_bookings.csv dengan format .csv. pada penelitian ini data akan di oleh dengan menggunakan bahasa pemograman *python*.

2. Menghapus fitur

Setelah data sudah dibaca dilakukan penghapusan fitur-fitur yang dianggap tidak berkaitan dengan target yang ingin dicapai. Fitur yang dihapus adalah "reservation_status" karena memiliki nilai yang sama dengan "isCancelled" dimana jika keduanya dipakai maka akan menghasilkan nilai akurasi 100% yang mana tidak ideal. Pada fitur "reservation_status_date" tidak digunakan karena tidak ada korelasinya pada setiap fitur.

3. Menghilangkan *Missing Value*

Lalu selanjutnya dilakukan imputasi *missing value* yang ada pada beberapa fitur dan disesuaikan dengan jenis tipe dari fitur tersebut, misal fitur yang berisi data kategorikal akan diisi menggunakan nilai yang paling sering muncul, dan fitur bertipe data numerik menggunakan nilai 0.

4. Implementasi *LabelEncoder*, Pengurutan Data, Deklarasi Kelas

Seluruh fitur yang berisi tipe kategorikal kemudian dilabeli menggunakan LabelEncoder. Kemudian mengurutkan data berdasarkan letak hotel, nilai 0 untuk Resort _hotel dan nilai 1 untuk City_hotel. Setelah itu dideklarasikan target atau kelas yang dituju, dalam data ini target nya adalah fitur "is_canceled".

5. Pembagian data

Selanjutnya dilakukan pembagian data. Data yang digunakan akan dibagi menjadi training dan testing. Dalam pembagian data metode yang digunakan untuk membagi data adalah metode Kfold dengan membagi data sebanyak 5-*fold*.

6. Pemodelan dan Hasil Akurasi data dengan *Decision Tree* dan *Random Forest*

Setelah data dibagi maka selanjutnya dilakukan pemodelan dengan menggunakan data latih atau data traning. Pada penelitian ini model yang digunakan *Decision Tree* dan *Random Forest*. Setelah terbentuk sebuah model data akan diuji dengan menggunakan data testing, Sehingga didapatkan nilai akurasi dari hasil prediksi dengan menggunakan kedua model yaitu model *decision tree* dan *random forest*.

7. Seleksi Fitur

Langkah selanjutnya adalah melakukan seleksi fitur untuk melihat fitur mana saja yang lebih berpengaruh dibanding fitur lainnya. Pada penelitian ini seksi fitur dilakukan dengan menggunakan metode PCA yang merupakan teknik yang digunakan untuk menyederhanakan suatu data, dengan cara mentransformasi data secara linier sehingga terbentuk sistem koordinat baru dengan varian maksimum. Setelah dilakukan pembobotan pada

setiap fitur dilakukanya pengurutan bobot fitur menggunakan fitur yang sudah disederhanakan tersebut, dalam penelitian Ini menggunakan perbandingan 3, 5 dan 14 fitur. Setelah itu dilakukan pemodelan kembali menggunakan model yang sama dengan fitur yang sudah disederhanakan, kemudian dicari kembali nilai akurasinya lalu dilihat apakah nilai akurasi yang didapat lebih baik dari menggunakan seluruh fitur.

4 Hasil dan Pembahasan

Pada bab ini akan membahas hasil penelitian tentang pengaruh seleksi fitur pada algoritma *machine learning*, seperti sudah dijelaskan pada bab sebelumnya bahwa algoritma yang dipakai pada penelitian ini adalah algoritma *decision tree* dan *random forest* yang mana setelahnya akan diaplikasikan seleksi fitur menggunakan PCA (*principal component analysis*) dalam melakukan seleksi fitur yang nantinya akan dilakukan prediksi dan akan dibandingkan berdasarkan hasil akurasi prediksinya dengan penerapan algoritma *machine learning* yang sama tetapi tanpa menggunakan PCA. Dengan menggunakan data yang didapat dari penelitian yang sudah dilakukan sebelumnya oleh (Antonio, De Almeida, & Nunes, 2017) berupa data pemesanan hotel di Algarve, Portugal dari tahun 2015 hingga tahun 2017 dengan jumlah data berjumlah 119.390 data yang dibagi berdasarkan jenis hotel yaitu *resort hotel* dan *city hotel*. Pembagian data berdasarkan jenis hotel dapat dilihat pada gambar 1.

Gambar. 1. Pembagian Data Berdasarkan Jenis hotel

Pada gambar 1 dapat dilihat bahwa jumlah data pada *city hotel* lebih banyak dibandingkan data yang termasuk kedalam *resort hotel*, lalu dari kedua jenis hotel tersebut akan dilihat banyaknya pembatalan yang terjadi pada kedua jenis hotel tersebut yang dapat dilihat pada gambar 2.

Gambar. 2. Jumlah Pembatalan Pesanan Hotel Pada Kedua Jenis Hotel

Jumlah pembatalan pesanan hotel seperti yang diilustrasikan oleh gambar 2 bahwa pada *resort hotel* mengalami peristiwa pembatalan pesanan hotel sebesar 27,76% dari total pesanan untuk jenis hotel tersebut, sementara pada *city hotel* terjadi pembatalan pesanan hotel sebesar 41,72% dari total pesanan untuk jenis hotel tersebut. Hal ini menunjukkan bahwa pada *city hotel* terjadi pembatalan pesanan terbanyak tetapi juga memiliki jumlah pesanan terbanyak yang tidak dibatalkan.

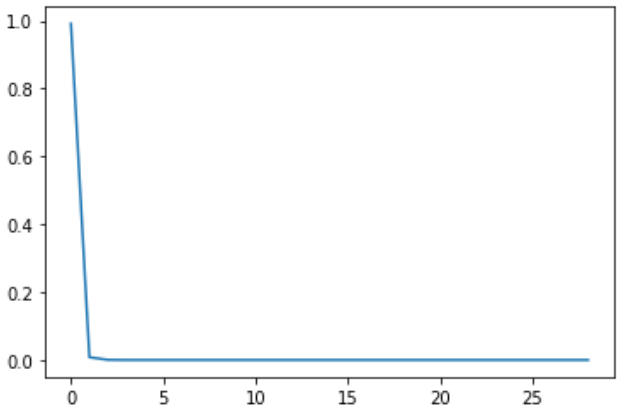
Kemudian, data akan dibagi menjadi 5 bagian yang ukurannya relatf sama menggunakan *k-fold* dimana 1 bagian data akan bertindak sebagai data uji, dan 4 bagian data sisanya akan bertindak sebagai data latih. Setelah dibagi menjadi data latih dan data uji, maka dapat dilakukan prediksi menggunakan algoritma *decision tree* dan *random forest* tanpa menggunakan PCA dengan hasil sebagai berikut :

Tabel 1. Hasil Akurasi dengan *Decision Tree* dan *Random Forest*

Model	**Akurasi**	***Precission***	***Recall***	***f1-Score***
Decision Tree	0.982	0.98	0.98	0.98
Random Forest	0.987	0.99	0.99	0.99

Hasil yang didapat menunjukkan bahwa pemodelan menggunakan algoritma *random forest* memiliki nilai akurasi yang lebih tinggi dari pemodelan menggunakan *decision tree*.

Karena tujuan dari penelitian ini adalah membandingkan hasil akurasi Ketika menggunakan PCA dan tanpa menggunakan PCA, maka selanjutnya dilakukan pemeringkatan fitur mulai yang dapat dilihat pada gambar 3.



Gambar. 3. Grafik Peringkat fitur

Pada grafik tersebut dapat dilihat bahwa dari peringkat fitur ke-1 memiliki nilai yang signifikan dalam mempengaruhi hasil prediksi. Setelah melewati peringkat fitur ke-5 nilainya cenderung melandai. Bobot fitur dalam bentuk angka dapat dilihat pada tabel 2.

Tabel 2. Bobot Fitur

Index	**Bobot Fitur Data Hotel**	**Index**	**Bobot Fitur Data Hotel**	**Index**	**Bobot Fitur Data Hotel**
0	9.919255e+01	13	7.897816e-08	26	3.573727e-09
1	8.060628e-01	14	7.061380e-08	27	1.660112e-09
2	9.841068e-04	15	5.532978e-08	28	6.769565e-10
3	2.082100e-04	16	4.289540e-08		
4	1.494601e-04	17	3.732893e-08		
5	2.617422e-05	18	2.949250e-08		
6	1.420783e-05	19	2.451310e-08		
7	6.246361e-06	20	2.334391e-08		
8	9.054705e-07	21	2.223049e-08		
9	4.122487e-07	22	1.675466e-08		

e-ISBN 978-623-93343-3-8
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 20 April 2021

10	2.697113e-07	23	1.146124e-08
11	1.549226e-07	24	9.636512e-09
12	1.052117e-07	25	5.932784e-09

Kemudian, dari hasil pemeringkatan fitur tersebut dilakukan prediksi dengan seleksi fitur menggunakan PCA dengan 3,5, dan 14 fitur karena berdasarkan pengamatan, dari peringkat ke-3 sampai ke-5 masih terdapat perubahan yang cukup besar sedangkan setelah melewati peringkat tersebut perubahannya tidak signifikan. Prediksi dilakukan menggunakan algoritma yang sama dengan sebelumnya, hasil prediksi menggunakan dengan 3,5, dan 14 fitur dan perbandingannya dengan hasil prediksi tanpa seleksi fitur dapat dilihat pada tabel 3.

Tabel 3. Perbandingan Hasil Akurasi *Decision Tree* dan *Random Forest* Menggunakan Seleksi Fitur dan Tanpa Seleksi Fitur

Model	**TP**	**TN**	**FP**	**FN**	**Akurasi**	***Precission***	***Recall***	***f1-score***
Decision Tree	14746	14731	298	291	0.980	0.98	0.98	0.98
Random Forest	14859	14831	198	178	0.987	0.99	0.99	0.99
Decision Tree dengan 3 Fitur	14333	14359	670	704	0.954	0.96	0.95	0.95
Random Forest dengan 3 Fitur	14440	14517	512	597	0.963	0.97	0.96	0.96
Decision Tree dengan 5 Fitur	14689	14649	380	348	0.966	0.97	0.96	0.97
Random Forest dengan 5 Fitur	14499	14556	473	538	0.976	0.97	0.98	0.98
Decision Tree dengan 14 Fitur	14743	14791	238	294	0,971	0.97	0.97	0.97
Random Forest dengan 14 Fitur	14601	14651	378	436	0,980	0.98	0.98	0.98

Dari tabel 3 dapat dilihat bahwa nilai akurasi pada algoritma *decision tree* dengan menggunakan 3,5, dan 14 fitur berturut-turut adalah 0,954, 0,963, dan 0,971. Sementara nilai akurasi algoritma *random forest* dengan menggunakan 3,5, dan 14 fitur berturur-turut adalah 0,963, 0,976, 0,980. Hal ini menunjukkan bahwa penggunaan PCA untuk melakukan seleksi fitur justru mengurangi akurasi dari prediksi yang dilakukan dan prediksi tanpa menggunakan PCA memiliki hasil yang lebih baik.

5 Kesimpulan dan Saran

Penggunaan PCA untuk melakukan seleksi fitur pada penelitian ini menghasilkan nilai yang meningkat seiring dengan bertambahnya fitur yang dipakai, tetapi setelah 14 fitur digunakan nilai akurasi yang didapat tidak mampu melebihi ataupun menyamai nilai akurasi pada data yang tidak menggunakan seleksi fitur. Dalam kasus ini PCA tidak mampu menjadi teknik seleksi fitur tunggal dikarenakan PCA memiliki beberapa kelemahan yang tidak bisa ditangani.

Penelitian selanjutnya diharapkan dapat memperbaiki hasil yang didapat pada penelitian ini dengan menggunakan dengan mengombinasikan PCA dengan teknik lainnya pada seleksi fitur untuk mengatasi kelemahan yang dimiliki oleh PCA.

Referensi

[1] S. F. Putra, R. Pradina, and I. Hafidz, “Feature Selection pada Dataset Faktor Kesiapan Bencana pada Provinsi di Indonesia Menggunakan Metode PCA (Principal Component Analysis),” *J. Tek. ITS*, vol. 5, no. 2, 2016, doi: 10.12962/j23373539.v5i2.16035.

[2] D. Jain and V. Singh, “An Efficient Hybrid Feature Selection model for Dimensionality Reduction,” *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 333–341, 2018, doi: 10.1016/j.procs.2018.05.188.

[3] R. Adhao and V. Pachghare, “Feature selection using principal component analysis and genetic algorithm,” *J. Discret. Math. Sci. Cryptogr.*, vol. 23, no. 2, pp. 595–602, 2020, doi: 10.1080/09720529.2020.1729507.

[4] Z. Wu, W. Lin, Z. Zhang, A. Wen, and L. Lin, “An Ensemble Random Forest Algorithm for Insurance Big Data Analysis,” *Proc. - 2017 IEEE Int. Conf. Comput. Sci. Eng. IEEE/IFIP Int. Conf. Embed. Ubiquitous Comput. CSE EUC 2017*, vol. 1, pp. 531–536, 2017, doi: 10.1109/CSE-EUC.2017.99.

[5] Suyanto, *Machine Learning Tingkat Dasar dan Lanjut*. Bandung: Informatika Bandung, 2018.

[6] A. S. Khazari, F. Marisa, and I. D. Wijaya, “Sistem Rekomendasi Penentuan Judul Skripsi Menggunakan Algoritma Decision Tree,” *J. Teknol. dan Manaj. Inform.*, vol. 3, no. 1, 2017, doi: 10.26905/jtmi.v3i1.1248.

[7] N. Antonio, A. de Almeida, and L. Nunes, “Hotel booking demand datasets,” *Data Br.*, vol. 22, pp. 41–49, 2019, doi: 10.1016/j.dib.2018.11.126.

[8] N. António, A. de Almeida, and L. Nunes, “Predictive models for hotel booking cancellation: a semi-automated analysis of the literature,” *Tour. Manag. Stud.*, vol. 15, no. 1, pp. 7–21, 2019, doi: 10.18089/tms.2019.15011.