PENGARUH SELEKSI FITUR PADA ALGORITMA *MACHINE LEARNING* UNTUK MEMPREDIKSI PEMBATALAN PESANAN HOTEL

Nama Anggota:

Farhana Nabila	1710511006
Siti Antania Syifa	1710511017
Imha Luchman Hakim	1710511040
Hersa Magdalena De Win	1710511061
I Gusti Naufhal Daffa Adnyana	1710511082

PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAKARTA

2020

Abstrak

Perkembangan penggunaan internet yang pesat saat ini, menunjukkan adanya pergesaran teknologi yang semakin maju mengarah ke media berbasis online. Konsumen cenderung untuk menelusuri (surfing) kelengkapan informasi produk dan jasa melalui internet dan melakukan pembelian secara online dikarenakan keterbatasan waktu serta kemudahan yang dirasakan. Penelitian yang dilakukan bertujuan untuk untuk menunjukkan bagaimana ilmu data dapat diterapkan di Internet pada konteks manajemen pendapatan hotel untuk memprediksi pembatalan pesanan hotel menggunakan model *decision tree* dan *random forest* lalu menggunakan atau PCA sebagai alat seleksi fitur untuk menemukan fitur mana yang lebih berpengaruh. Rancangan penelitian yang digunakan bertujuan untuk menguji hipotesis yang dikembangkan berdasarkan teori-teori terkait. Data dianalis dengan metode *decision tree* . Hasil penelitian menemukan bahwa: Pembatalan pemesanan memiliki dampak besar dalam keputusan manajemen permintaan di industri perhotelan. Pembatalan membatasi produksi perkiraan akurat, alat penting dalam hal kinerja manajemen pendapatan. Untuk menghindari masalah yang disebabkan oleh pembatalan pemesanan, hotel menerapkan kebijakan pembatalan yang kaku dan strategi *overbooking*, yang juga dapat memiliki pengaruh negatif pada pendapatan dan reputasi. Menggunakan set data dari empat hotel yang dikelompokkan kedalam jenis hotel resor dan hotel kota, peneliti menunjukkan bahwa dimungkinkan untuk membangun model untuk memprediksi pembatalan pemesanan dengan hasil akurasi lebih dari 90%. Hasil ini memungkinkan manajer hotel untuk secara akurat memprediksi permintaan bersih dan membuat perkiraan yang lebih baik, memperketat kebijakan pembatalan, menentukan taktik menghadapi *overbooking* yang lebih baik dan dengan demikian menggunakan strategi penetapan harga dan alokasi inventaris yang lebih tegas.

Kata Kunci: *random forest, decision tree*, PCA

1. Pendahuluan

Manajemen pendapatan didefinisikan sebagai “aplikasi dari sistem informasi dan strategi penetapan harga untuk mengalokasikan hak kepada pelanggan yang tepat, harga yang tepat, dan waktu yang tepat” (Kimes & Wirtz, 2003, hal. 125). Awalnya dikembangkan pada tahun 1966 oleh industri penerbangan (Chiang, Chen, & Xu, 2007). Pemesanan merupakan kontrak antara pelanggan dan hotel (Talluri & Van Ryzin, 2004). Kontrak ini memberi pelanggan hak untuk menggunakan layanan di masa depan dengan harga tetap, biasanya dengan opsi untuk membatalkan kontrak sebelum ketentuan layanan. Meskipun pemesanan lanjut dianggap sebagai yang terkemuka alat prediksi kinerja perkiraan hotel (Smith, Parsa, Bujusuc, & van der Rest, 2015), opsi ini untuk membatalkan layanan menempatkan risiko pada hotel. Pembatalan terjadi ketika pelanggan mengakhiri kontrak sebelum kedatangannya dan tidak muncul terjadi ketika pelanggan tidak menginformasikan hotel dan gagal check-in, atau terjadi karena pelanggan yang mencari kesepakatan yang ditentukan dalam mencari penawaran terbaik. Memang, reservasi dengan opsi pembatalan memberi pelanggan yang terbaik dari kedua dunia — manfaat dari mengunci ketersediaan maju dan fleksibilitas untuk mengingkari apakah rencana mereka atau preferensi berubah ”.

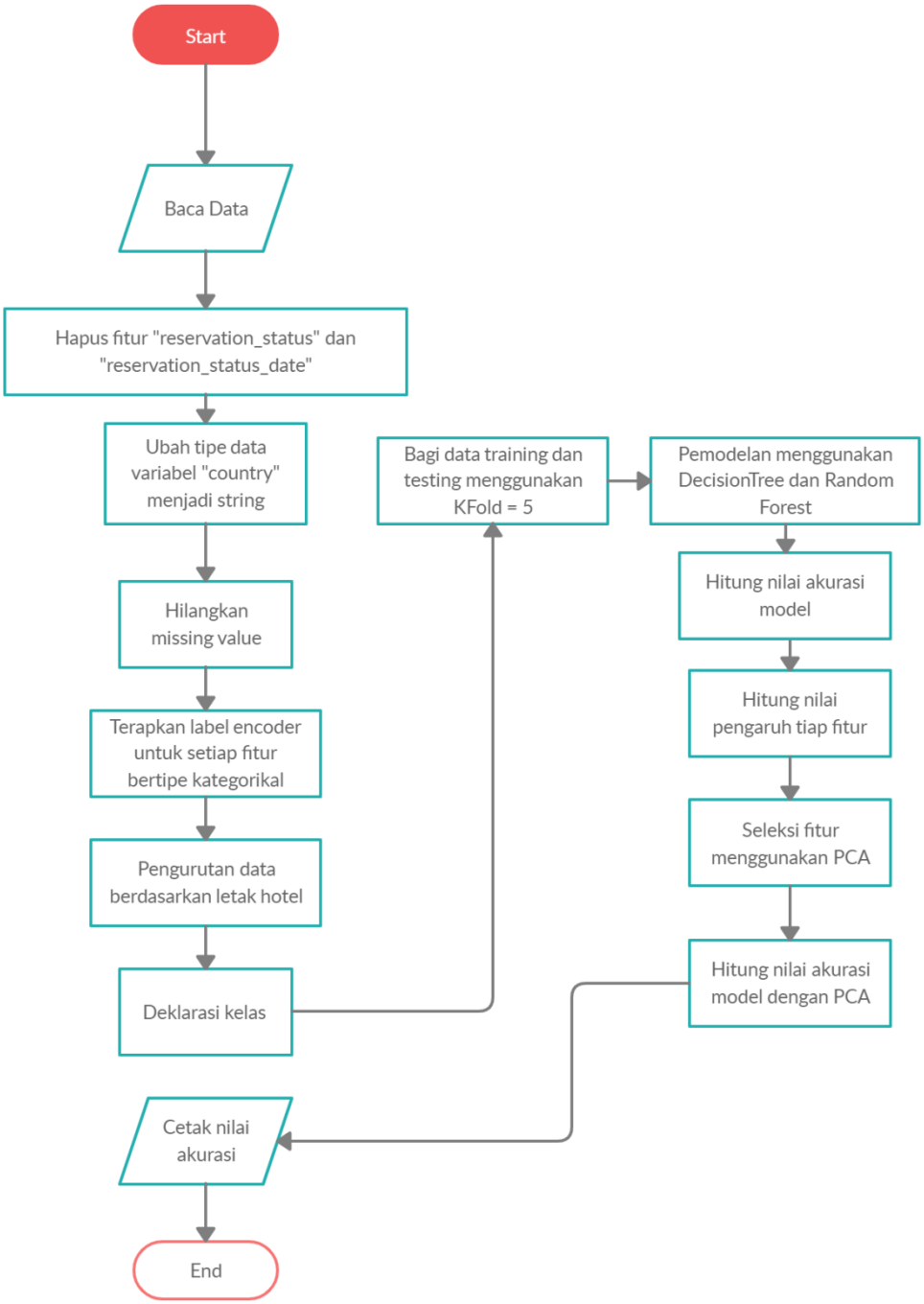
Sebagai cara untuk mengelola risiko yang terkait dengan pembatalan pemesanan,hotel menerapkan kombinasi overbooking dan pembatalan kebijakan (C.-C. Chen et al., 2011; C.-C. Chen & Xie, 2013; Mehrotra & Ruttley, 2006; Smith et al., 2015; Talluri & Van Ryzin, 2004). Namun, kebijakan pemindahbukuan dan pembatalan dapat dilakukan merugikan hotel. Overbooking, dengan tidak mengizinkan pelanggan untuk check-in di hotel yang sebelumnya dipesan, memaksa hotel untuk menolak penyediaan layanan kepada pelanggan, yang bisa menjadi pengalaman yang mengerikan bagi pelanggan. Pengalaman ini dapat memiliki efek negatif pada reputasi hotel dan pendapatan langsung (Noone & Lee, 2011), belum lagi potensi hilangnya pendapatan di masa mendatang dari pelanggan yang tidak puas tidak akan memesan lagi untuk menginap di hotel (Mehrotra & Ruttley, 2006).

Studi ini menunjukkan bahwa manajemen pendapatan akan semakin strategis dan berorientasi teknologi dan profesional manajemen pendapatan harus memiliki yang lebih baik dalam keterampilan analitis dan komunikasi. Studi ini mengidentifikasi semua profesional manajemen pendapatan harus memiliki analitis dan keterampilan komunikasi, yang merupakan keterampilan yang tepat pada akar ilmu data. Sebagai disiplin yang cukup baru, ilmu data mengambil keuntungan dari sejumlah besar data yang kami miliki dan ketersediaan daya komputasi yang lebih baik dan lebih murah. Ini faktor-faktor yang memungkinkan peningkatan prediksi yang ada algoritma dan berkontribusi pada pengembangan baru dan algoritma yang lebih baik, khususnya di bidang pembelajaran mesin.

Tugas ini bertujuan untuk menunjukkan bagaimana ilmu data dapat diterapkan di Internet dalam konteks manajemen pendapatan hotel untuk memprediksi pembatalan pesanan hotel. Pada penelitian sebelumnya digunakan 5 metode untuk mandapatkan nilai akurasi, yaitu *boosted decision tree, decision jungle, locally deep support vector machine, dan neural network*. Pada penelitian ini kami menggunakan model *decision tree* dan *random forest* untuk kemudian dilihat nilai akurasinya, lalu dilakukan seleksi fitu menggunakan PCA untuk menemukan mana fitur yang lebih berpengaruh. Setelah dilakukan seleksi fitur menggunakan PCA, data kemudian akan dimodelkan kembali menggunakan model sebelumnya lalu dilihat apakah nilai akurasi yang didapat bisa lebih baik dari sebelumnya atau tidak.

2. Metodologi Penelitian

2.1 Flow Chart



Pertama-tama, data akan dibaca terlebih dahulu oleh IDE python yaitu Spyder. Data yang digunakan adalah data hotel_bookings.csv, setelah data sudah dibaca dilakukan penghapusan fitur-fitur yang dianggap tidak berkaitan dengan target yang ingin dicapai, fitur yang dihapus adalah "reservation_status_date" dan "reservation_status" dimana jika keduanya dipakai maka akan menghasilkan nilai akurasi 100% yang mana tidak ideal.

Pada data yang digunakan, pada fitur "country" berisi nama-nama negara sehingga dianggap sebagai tipe data string agar dapat diolah oleh sistem. 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.

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".

Selanjutnya dilakukan pembagian data untuk training dan testing, metode yang digunakan untuk membagi data adalah dengan metode Kfold dengan membagi data sebanyak 5-*fold*. Setelah data dibagi maka selanjutnya dilakukan pemodelan, pada percobaan ini kami menggunakan model *decision tree* dan *random forest*, setelah itu didapatkan nilai akurasinya untuk data hotel keseluruhan, resort hotel, dan city hotel.

Langkah selanjutnya adalah melakukan seleksi fitur untuk melihat fitur mana saja yang lebih berpengaruh dibanding fitur lainnya, dalam melakukan hal ini kami menggunakan metode PCA yang merupakan teknik yang digunakan untuk menyederhanakan suatu data, dengan cara mentransformasi data secara linier sehingga terbentuk sistem koordinat baru dengan varians maksimum. Setelah fitur-fitur diperingkat, kami menggunakan fitur yang sudah disederhanakan tersebut, dalam hal ini kami menggunakan 4 fitur saja. Setelah itu dilakukan pemodelan kembali menggunakan model yang sama tapi dengan fitur yang sudah disederhanakan, kemudian dicari kembali nilai akurasinya lalu dilihat apakah nilai akurasi yang didapat lebih baik dari menggunakan seluruh fitur.

3. Hasil dan Pembahasan

Berdasarkan langkah-langkah yang sudah dijelaskan sebelumnya yaitu penerapan metode klasifikasi decision tree dan random forest pada dataset hotel secara umum, dataset resort hotel, dan dataset city hotel tanpa melakukan seleksi fitur dan dengan menggunakan seleksi fitur menggunakan PCA, dilakukan proses analisis berdasarkan nilai akurasi yang didapatkan pada proses klasifikasi, analisis ini juga bertujuan untuk mengetahui apakah klasifikasi setelah melakukan seleksi fitur dengan menggunakan PCA dapat menghasilkan nilai akurasi yang lebih baik daripada menggunakan keseluruhan fitur. Nilai bobot tiap fitur dapat dilihat pada Tabel 1, adapun hasil pengujian dari ketiga dataset dapat dilihat sebagai berikut:

Tabel 2. Hasil Akurasi

Tabel 1. Peringkat Fitur

Index	Bobot fitur data hotel	Bobot fitur data Resort Hotel	Bobot fitur data City Hotel
0	990607.7993	0.994103787	0.976464876
1	9378.707006	0.005882382	0.023512839
2	9.462323109	1.00E-05	1.38E-05
3	1.992864733	1.84E-06	5.32E-06
4	1.560424198	1.40E-06	2.68E-06
5	0.239365603	3.36E-07	2.79E-07
6	0.150007205	1.35E-07	1.16E-07
7	0.064583371	6.26E-08	7.77E-08
8	0.009283882	9.19E-09	1.48E-08
9	0.004392652	2.90E-09	1.00E-08
10	0.002904442	2.27E-09	8.22E-09
11	0.00185459	1.75E-09	2.74E-09
12	0.001170295	1.14E-09	2.49E-09
13	0.000854068	7.74E-10	1.44E-09
14	0.000591089	5.91E-10	1.22E-09
15	0.000577262	4.49E-10	1.09E-09
16	0.000479905	3.61E-10	8.76E-10
17	0.000451418	2.92E-10	8.34E-10
18	0.000345859	2.40E-10	7.32E-10
19	0.000253438	2.30E-10	6.22E-10
20	0.000247172	1.74E-10	4.67E-10
21	0.000209832	1.38E-10	4.01E-10
22	0.000171042	8.96E-11	2.90E-10
23	0.000118644	8.74E-11	2.33E-10
24	0.000104479	5.01E-11	1.73E-10
25	5.43E-05	1.94E-11	6.95E-11
26	4.56E-05	1.41E-11	4.66E-11
27	1.92E-05	4.93E-12	2.07E-11
28	7.43E-06		

Data	Model	Akurasi Tanpa Seleksi Fitur	Akurasi Dengan menggunakan 3 Fitur	Akurasi Dengan menggunakan 5 Fitur	Akurasi Dengan menggunakan 14 Fitur
Hotel Secara Umum	Decision Tree	0.976	0.958	0.966	0.966
	Random Forest	0.982	0.966	0.974	0.974
Resort Hotel	Decision Tree	0.976	0.965	0.970	0.968
	Random Forest	0.983	0.972	0.976	0.978
City Hotel	Decision Tree	0.972	0.946	0.969	0.968
	Random Forest	0.982	0.956	0.975	0.975

Dari Tabel 2 terlihat bahwa nilai akurasi dengan menggunakan seluruh fitur lebih tinggi daripada nilai akurasi dengan seleksi fitur menggunakan PCA walaupun hanya berbeda ±3%. Klasifikasi dengan *decision tree* pada data hotel secara umum tanpa menggunakan PCA memiliki nilai akurasi sebesar 0.976 dari skala 1, sementara jika menggunakan PCA dengan 3,5,14 fitur berturut-turut adalah sebesar 0.958, 0.966, dan 0.966. Sementara untuk data yang sama dengan menggunakan *random firest* menghasilkan nilai akurasi berturut-turut sebesar 0.982,0.966,.0.974, dan 0.974. Untuk data *Resort Hotel* dengan *decision tree* mendapatkan nilai akurasi berturut-turut sebesar 0.976,0.965,0.970,0.968, disini terlihat pada PCA dengan 5 fitur dengan PCA dengan 14 fitur terjadi penurunan nilai akurasi. Jika menggunakan *random forest* pada data yang sama, nilai akurasinya berturut-turut adalah 0.983,0.973,0.976,0.978, terjadi kenaikan nilai akurasi seiring bertambahnya fitur yang digunakan. Pada data *City Hotel* dengan menggunakan *decision tree* menghasilkan nilai akurasi sebesar 0.972,0.946,0.969,0.968, disini terjadi penurunan kembali seperti pada data *Resort Hotel* pada PCA dengan 14 fitur. Semestara jika menggunakan *random forest*, nilai akurasi yang didapat berturut-turut sebesar, 0.982,0.956,0.975,0.975 pada data ini terjadi stagnan nilai akurasi pada PCA 5 dan PCA 14 dimana keduanya menghasilkan nilai akurasi yang sama. Sehingga dapat ditarik kesimpulan bahwa model tanpa seleksi fitur dengan PCA memiliki nilai akurasi yang lebih tinggi dibanding menggunakan PCA, sementara itu metode klasifikasi dengan *random forest* memiliki kinerja yang lebih baik dibangingkan dengan metode *decision tree*.

4. 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.

Secara umum metode klasifikasi dengan *random forest* memiliki kinerja yang lebih baik daripada metode klasifikasi dengan *decision tree* baik menggunakan PCA ataupun tidak. Hal tersebut dapat dilihat dari nilai akurasi yang dihasilkan oleh kedua metode.

Penelitian selanjutnya diharapkan dapat memperbaiki hasil yang didapat pada penelitian ini dengan melakukan praproses data yang lebih baik atau memilih metode klasifikasi yang lain serta mampu mengimplementasikan sistem ini pada data yang berbeda.

5. Daftar Pustaka

Antonio, N., de Almeida, A. and Nunes, L., (2019). Predicting Hotel Booking Cancellation to Decrease Uncertainty and Increase Revenue. *Tourism & Management Studies*, 13(2), pp.25-39.

Antonio, N., de Almeida, A. and Nunes, L., (2019). Hotel booking demand datasets. *Data in brief*, 22, pp.41-49.

Kimes, S. E., & Wirtz, J. (2003). Has revenue management become acceptable? Findings from an International Study on the Perceived Fairness of Rate Fences. *Journal of Service Research, 6*(2), 125-135.

Chiang, W.-C & Chen, J.C.H. & Xu, X.. (2007). An Overview of Research on Revenue Management: Current Issues and Future Research. 1. 97-128.

Talluri, Kalyan & van Ryzin, Garrett. (2004). The Theory and Practice of Revenue Management. 10.1007/b139000.

Parsa, H. & van der Rest, Jean-Pierre & Smith, Scott & Parsa, Rahul & Bujisic, Milos. (2015). Why Restaurants Fail? Part IV: The Relationship between Restaurant Failures and Demographic Factors. Cornell Hospitality Quarterly. 56. 80-90.