E-ISSN 2808-5841
P-ISSN 2808-7283

Sistem Rekomendasi Pencarian Indekos di Surabaya Menggunakan Random Forest

Theresa Agnes Virnauli Sinaga[1], Lucia Bellanie Debra[2], Jasmine Taj Ariva[3], Devi Rizky Aditya[4], Vivia Faustine Gunawan[5], Maryamah Maryamah[6]

[1,2,3,4,5] *Teknologi Sains Data, Universitas Airlangga*

[1] theresa.agnes.virnauli-2021@ftmm.unair.ac.id

[2] lucia.bellanie.debra-2021@ftmm,unair.ac.id

[3] jasmine.taj.ariva-2021@ftmm.unair.ac.id

[4] devi.rizky.aditya-2021@ftmm.unair.ac.id

[5] vivia.faustine.gunawan-2021@ftmm.unair.ac.id

[6] maryamah@ftmm.unair.ac.id

Corresponding author email: maryamah@ftmm.unair.ac.id

Abstract: *Looking for a boarding house is common for college students as a temporary place to live when they are studying outside their area of residence. Obtaining information regarding boarding prices that meet the criteria desired by students is quite difficult. In this paper, we proposed a recommendation system for finding boarding houses according to the criteria desired by students using the random forest method. This system can help students get boarding prices that match the boarding criteria they want, especially in the Surabaya area. The research method starts with data collection, preprocessing, and model training using the random forest. Based on the experimental results using the Decision Tree and Support Vector Machine (SVM) comparison method, the proposed method has the highest accuracy rate with a value of 78.55% and an error rate of 305887.80º. This recommendation system for prediction boarding house can help and make it easier for students to find boarding houses that match the criteria they want.*

Keywords: *Recommendation systems, boarding house, Machine Learning, random forest*

Abstrak: Mencari tempat kos merupakan hal yang umum bagi mahasiswa sebagai tempat tinggal sementara ketika mereka menempuh masa pendidikan di luar daerah tempat tinggal mereka. Mendapatkan informasi terkait harga kos yang sesuai kriteria yang diinginkan oleh mahasiswa adalah hal yang cukup sulit. Oleh karena itu, kami mengusulkan sistem rekomendasi pencarian rumah kos sesuai kriteria yang diinginkan mahasiswa menggunakan metode *random forest*. Sistem ini dapat membantu mahasiswa mendapatkan harga kos yang sesuai dengan kriteria kos yang mereka inginkan khususnya di wilayah Surabaya. Metode penelitian dimulai dari pengumpulan data, preprocessing, pembuatan model menggunakan algoritma *random forest*. Berdasarkan hasil eksperimen dengan metode perbandingan *Decision Tree* dan *Support Vector Machine* (SVM), metode yang diusulkan memiliki tingkat akurasi paling tinggi dengan nilai sebesar 78.55% dan tingkat error sebesar 305887.80**º.** Prediksi harga ini dapat membantu dan mempermudah mahasiswa dalam pencarian kos yang sesuai dengan kriteria kos yang mereka inginkan.

Kata kunci: Sistem rekomendasi, Pencarian kos, *Machine Learning*, *random forest*

I. PENDAHULUAN

Menurut Badan Pusat Statistik (BPS) Provinsi Jawa Timur Kota Surabaya pada tahun 2020, jumlah perguruan tinggi terhitung sebanyak 76 dengan total mahasiswa sejumlah 257.630 orang [1]. Berdasarkan data yang diperoleh dari survei pribadi peneliti terhadap responden yang berstatus sebagai mahasiswa di wilayah Surabaya, menunjukkan bahwa dari 54 responden hanya satu responden yang tidak memilih kos sebagai opsi tempat tinggal selama menempuh pendidikan. Dengan menyesuaikan kebutuhan mahasiswa yang membutuhkan tempat tinggal, keberadaan kos tentu memudahkan para mahasiswa khususnya dari luar daerah tempat mereka menuntut ilmu di lembaga pendidikan terkait. Kos atau *boarding house* didefinisikan sebagai tempat tinggal sementara yang disewakan oleh pemilik atau pengelola dengan harga terjangkau kepada mahasiswa atau pekerja. Kos umumnya terdiri dari beberapa kamar yang disewakan secara individu dengan dilengkapi fasilitas dasar seperti kamar mandi,

tempat tidur, lemari, meja, dan kursi. Kos umumnya juga menyediakan fasilitas tambahan seperti dapur bersama, ruang tamu, atau akses internet [2].

Ada banyak opsi hunian yang dapat dipertimbangkan oleh mahasiswa diantaranya rumah kontrakan, kos, rumah susun, dan masih banyak lagi yang disesuaikan dengan kenyamanan serta kemampuan dari calon mahasiswa. Namun, berdasarkan survei peneliti terhadap mahasiswa Surabaya menunjukkan bahwa mereka cenderung memilih kos sebagai hunian sementara. Hal ini didasari oleh beberapa hal diantaranya jarak kos ke kampus yang dekat, fasilitas yang memadai berupa ada tidaknya kamar mandi dalam, ada tidaknya biaya administrasi tambahan, luas kamar, ketersediaan kasur, dan ketersediaan luas parkiran.

Dalam membantu mahasiswa dalam melakukan pemilihan kos terdapat beberapa *expert system berupa* recommendation system melalui penelitian yang telah dikembangkan. Salah satunya yaitu penelitian mengenai *web* sebagai sistem rekomendasi kos di Bali menggunakan metode *naive bayes* yang ditambahkan fitur *google map*. *Naive bayes* digunakan dalam pengembangan sistem rekomendasi ini karena dinilai memiliki akurasi yang tinggi [3]. Hasil diaplikasikan pada sistem dengan variabel input harga, jarak, kelengkapan fasilitas, yang kemudian menampilkan *output* rekomendasi kos berupa gambar kos, nama kos, dan harga. Selain itu, terdapat pula algoritma *Simple Additive Weighting* (*SAW*) pada penelitian rekomendasi tempat kos di Pringsewu dengan diterapkannya pembuatan *DSS* (*Decision Support System*). Metode tersebut berjalan berdasarkan penambahan bobot dari rating kinerja tiap alternatif pada atribut, serta menggunakan *decision matrix* normalisasi ke skala yang dapat dibandingkan dengan peringkat alternatif [1]. Hasil pengaplikasian disesuaikan dengan kriteria yang diinginkan sebagai *input*, dimana *output* berupa daftar kos yang layak.

Meskipun *recommendation system* kos telah dikembangkan, tetapi masih terdapat hal di dalam nya yang perlu ditingkatkan. Pertama, sistem rekomendasi belum memberikan prediksi harga terbaik berdasarkan fasilitas kos yang ditawarkan. Sehingga mahasiswa masih mengalami kesulitan dalam memperkirakan budget yang sesuai terhadap kriteria kos idaman yang mereka tetapkan. Mahasiswa butuh pemahaman pola harga kos dengan faktor penentunya agar dapat mencari tempat tinggal, berupa kos-kosan yang menarik sesuai kriteria yang dibutuhkan. Sehingga, diperlukannya sistem prediksi berbasis *machine learning* yang dapat membantu para mahasiswa memilih kos beserta fasilitas yang diperoleh, dimana penentuan tersebut berdasarkan informasi yang diberikan [4]. Selain itu, belum terdapat recommendation system yang secara khusus dikembangkan untuk mahasiswa di wilayah Surabaya.

Dari beberapa permasalahan tersebut, kami mengusulkan sistem rekomendasi pencarian rumah kos sesuai kriteria yang diinginkan mahasiswa menggunakan metode *machine learning random forest*. Sistem ini dapat membantu mahasiswa mendapatkan harga kos yang sesuai dengan kriteria kos yang mereka inginkan khususnya di wilayah Surabaya. Metode penelitian dimulai dari pengumpulan data, preprocessing, pembuatan model menggunakan algoritma random forest. Data yang digunakan adalah kombinasi data penelitian pribadi oleh peneliti dan *scraping* data dari website mamikos. Dengan penggunaan metode *machine learning* mampu menghasilkan akurasi yang lebih akurat dibandingkan beberapa algoritma lain.

II. METODE PENELITIAN

Metode penelitian yang digunakan dalam penulisan karya ilmiah ini dilandasi oleh beberapa referensi dari penelitian sebelumnya. Penelitian yang dikaji sebagai tinjauan pustaka adalah penelitian yang meneliti biaya sewa tempat tinggal dan aplikasi *expert system* berbasis *machine learning* sebagai pengambilan keputusan. Expert system merupakan sebuah sistem komputer untuk mengolah informasi yang ada menjadi sebuah penentu atau acuan dalam penentuan keputusan [5]. Penggunaan expert system dapat digunakan pada beberapa bidang seperti kesehatan untuk menganalisa diagnosis, pertanian untuk mengklasifikasikan hama, dan properti untuk menentukan harga sewa. Berdasarkan penelitian

independen yang dilakukan untuk memprediksi harga sewa rumah di Tanzania, Uganda dan Malawi pada tahun 2021 menggunakan *machine learning* memberikan hasil yang lebih baik dibandingkan menggunakan metode regresi Ordinary Least Square (OLS). Machine learning memperhitungkan dimensional struktur data dan korelasi antar variabel [6].

2.1 Teknik Pengambilan Data

Data yang digunakan pada penulisan karya ilmiah ini diambil dari dua teknik, yaitu data yang didapatkan melalui *web scraping* pada laman 'mamikos.com' sebagai data sekunder dan data yang didapatkan melalui survei sebagai data primer. Survei dilakukan dalam lingkup Universitas Airlangga, dimana mahasiswa angkatan 2021 prodi Teknologi Sains Data menjadi responden dari kuesioner yang dibagikan.

2.2 Data *Pre-Processing*

Data yang didapatkan dari kedua teknik yang telah disebutkan sebelumnya kemudian dibersihkan dan digabungkan. Namun, karena sumber data yang berbeda sehingga perlu dikonversi satuannya untuk menyelaraskan kedua *dataset*. Setelah data diselaraskan, data kemudian dijadikan satu untuk menjadi data bersih. Data yang telah digabungkan tersebut kemudian dilakukan categorical *encoding variable*, dimana teknik yang digunakan adalah *dummy encoding* atau mengubah variabel kategorik menjadi variabel *binary* atau dikenal sebagai variabel *dummy* [7]. Teknik ini dilakukan terhadap seluruh variabel independen agar dapat dimasukkan ke dalam model pada langkah selanjutnya.

2.3 Klasifikasi

Data yang telah dibersihkan kemudian dilatih menggunakan teknik *random forest classifier* guna memberikan *feedback* kepada pengguna apakah harga dan fasilitas yang ditawarkan kos sudah optimal atau diluar harga optimal. Random Forest merupakan sebuah algoritma machine learning yang termasuk dalam ensemble learning. Algoritma ini membangun beberapa (lebih dari dua) pohon keputusan secara acak, dengan setiap pohon keputusan menggunakan sampel acak dari data pelatihan dan subset acak dari atribut. Random Forest memiliki beberapa kelebihan antara lain dalam menangani data kompleks dan besar, toleransi terhadap overfitting, kemampuan mengatasi variabel tidak penting, dan memberikan perkiraan keakuratan model yang dihasilkan. Random Forest menghitung rata-rata dari hasil pada berbagai *decision tree* yang diaplikasikan pada berbagai subset set data untuk meningkatkan akurasi prediksi [4]. Data ini kemudian dibentuk menjadi bentuk model yang nantinya akan menerima nilai yang diberikan pengguna dan program akan mengembalikan dalam bentuk harga rentang untuk spesifikasi kos yang dimaksud. Random forest menggunakan rumus entropi sebagai penentu tingkat ketidakmurnian atribut dan nilai *information gain* seperti pada Persamaan 1 dan 2.

$$Entropy\ (Y)\ =\ -\Sigma_i p\ (c|Y)\ log_2 p\ (c|Y) \qquad (1)$$

Pada Persamaan (1), Y merupakan variabel acak dari suatu dataset, c merupakan nilai kelas atau label yang mungkin diambil oleh variabel acak Y, lalu pada $p\ (c|Y)$ merupakan probabilitas dari kemunculan kelas c pada variabel acak Y. Selanjutnya, dilakukan penjumlahan untuk semua nilai c yang mungkin terhadap nilai p $(c|Y)$ yang dikalikan dengan log_2 dari nilai p $(c|Y)$ dan dikalikan lagi dengan negatif satu.

$$Information\ Gain\ (Y,a)\ =\ Entropy\ (Y)\ -\ \Sigma_{veValues} \left|\frac{Yv}{Ya}\right| Entropy(Yv) \qquad (2)$$

Pada Persamaan (2), values (a) merupakan semua nilai yang mungkin dalam himpunan kasus a. Kemudian Yv adalah subkelas dari Y dengan kelas v yang berhubungan dengan kelas a, sedangkan Ya adalah semua nilai yang sesuai dengan a.

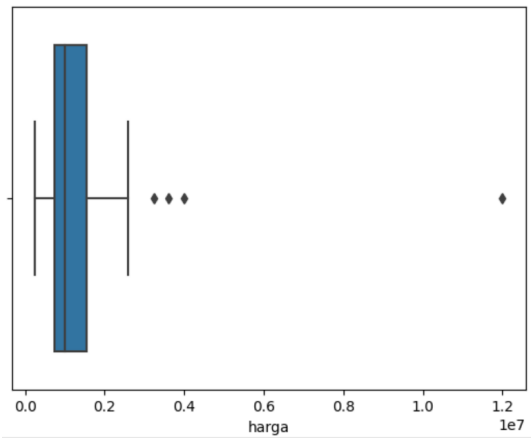
III. HASIL DAN PEMBAHASAN

Pada penelitian ini, data yang diperoleh akan diuji menggunakan tiga metode, yaitu *decision tree*, *random forest*, dan super *vector machine*. Untuk menguji metode mana yang paling baik, penelitian ini menggunakan skor uji akurasi dan *mean average error* (MAE). Metode yang paling baik adalah metode yang memiliki skor akurasi tertinggi dan nilai MAE terkecil. Sebelum membangun model, akan dilakukan analisis statistika deskriptif dari dataset dengan variabel-variabel bernama "harga", "leb-ih_satu_orang", "km_dalam", "wifi", "listrik", "ac", "tipe", yang mana ditampilkan hasil perhitungan rata-rata, standar deviasi, nilai minimum, dan nilai maksimum setelah data yang bersifat kategorik diubah menjadi *dummy variable*. Hasil tampilan analisis statistika deskriptif tersebut dapat pada gambar 1. Selanjutnya, mengecek nilai kosong atau NA guna untuk memastikan bahwasanya di dalam dataset tidak mengandung nilai yang kosong, dimana variabel dengan NA akan langsung dihapus. Dilakukannya modifikasi langsung pada dataset yang telah berbentuk data frame dengan menghapus baris duplikat.

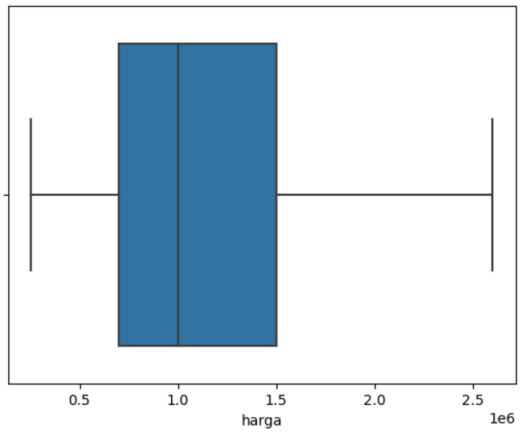
	harga	lebih_satu_orang	km_dalam	wifi	listrik	ac	tipe
count	1.010000e+02	101.000000	101.000000	101.000000	101.000000	101.000000	101.000000
mean	1.317495e+06	0.495050	0.485149	0.693069	0.554455	0.554455	1.297030
std	1.273151e+06	0.502469	0.502272	0.463521	0.499505	0.499505	0.806778
min	2.550000e+05	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	7.500000e+05	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
50%	1.000000e+06	0.000000	0.000000	1.000000	1.000000	1.000000	2.000000
75%	1.550000e+06	1.000000	1.000000	1.000000	1.000000	1.000000	2.000000
max	1.200000e+07	1.000000	1.000000	1.000000	1.000000	1.000000	2.000000

Gambar 1. Statistika Deskriptif Data

Selanjutnya untuk menentukan *outlier* yang perlu dihapus, dibuat visualisasi menggunakan *boxplot* untuk melihat harga kos yang *outlier*. Setelah nilai *outlier* diketahui, *outlier* kemudian dihapus menggunakan *lower bound* dan *upper bound* yang didapatkan dari nilai kuartil dan Interquartile Range (IQR). Perbedaan *boxplot* sebelum dan sesudah pembersihan dapat dilihat pada Gambar 2a. dan Gambar 2b. Setelah data dibersihkan, variabel x atau variabel prediktor dan variabel y atau variabel respon ditentukan, dimana variabel 'harga' menjadi variabel respon dan variabel lainnya menjadi variabel prediktor.



Gambar 2a. Variabel 'Harga' dengan outlier



Gambar 2b. Variabel 'Harga' setelah outlier dihapus

Perbandingan metode yang diusulkan, dibandingkan dengan dua metode *machine learning* lain yaitu *Decision Tree* dan *Support Vector Machine*. Nilai perbandingan dari ketiga metode

menggunakan hasil akurasi dan nilai *average error* pada Tabel 1. Nilai akurasi digunakan untuk melihat seberapa besar keakuratan model yang dihasilkan oleh algoritma. Pada tabel terlihat bahwa ketiga metode memiliki besaran nilai yang berbeda pada nilai akurasi Nilai akurasi tertinggi terdapat pada algoritma *Random Forest* sebesar 78,55 %. Hal ini menunjukkan bahwa keputusan yang dihasilkan oleh *Decision Tree* memberikan keakuratan yang tinggi. Setelah itu nilai akurasi tertinggi selanjutnya dimiliki oleh *Decision Tree* sebesar 78.50% dan *Super Vector Machine* sebesar 74.43%. Kemudian terdapat nilai *average error* yaitu nilai *error* yang mungkin dihasilkan dari perhitungan. Untuk itu, diperlukan nilai *error* terkecil yang dimiliki oleh *Random Forest* sebesar 305887.80⁰. Berdasarkan kedua indikator tersebut untuk menentukan algoritma yang dapat digunakan sebagai pengambilan keputusan model dipilih dengan nilai akurasi tertinggi dan *average error* terkecil yang dimiliki oleh *Random Forest* dengan selisih sebesar 0.05 lebih tinggi dibandingkan *Decision Tree*. Metode *Random Forest* memiliki hasil pengukuran lebih baik dalam peningkatan akurasi karena pemilihan pembangkitan simpul anak untuk setiap node dilakukan secara acak dan diakumulasikan hasil klasifikasi [7]. Melihat metode lain sebagai pembanding yaitu *Decision Tree* dan *Support Vector Machine* memiliki alur dan perhitungan yang berbeda dibandingkan dengan *Random Forest*.

Tabel 1. Nilai Hasil Perbandingan hasil

Algoritma	Nilai Akurasi (%)	Average Error(⁰)
Decision Tree	78.50	311400.48
Random Forest	78.55	305887.80
Super Vector Machine	74.43	315600.00

Decision Tree menggunakan struktur pohon untuk memodelkan keputusan berdasarkan atribut data yang dipunya. Adapun Setiap simpul mewakili keputusan atau tes pada atribut, dan cabang-cabang menggambarkan hasil keputusan. Pohon keputusan dapat digunakan untuk klasifikasi dan regresi, dan membantu interpretasi hubungan antara atribut dan hasil keputusan. SVM mengambil garis pemisah optimal (hyperplane) dengan margin maksimal tiap kelas data yang berbeda. SVM mampu menangani data linier maupun non-linier dengan transformasi non-linier. Kelebihan SVM adalah dalam mengoptimalkan fungsi margin maksimal, namun kurang efektif digunakan apabila dataset dalam jumlah besar.

IV. KESIMPULAN

Sistem rekomendasi harga kos dengan mengaplikasikan *machine learning random forest* memiliki tingkat akurasi lebih tinggi dibandingkan metode lainnya dengan akurasi sebesar 78.55% dan nilai *error* sebesar 305887.80⁰**.** Metode pembanding yang digunakan *Decision Tree* dan *Support Vector Machine*. Sistem rekomendasi yang dibangun akan membantu menyelesaikan permasalahan kebanyakan penyewa kos, dalam hal ini khususnya mahasiswa mendapatkan kos sesuai dengan kriteria yang diinginkan. Penggunaan akurasi dan nilai *average error* digunakan karena kedua indikator tersebut dinilai cukup untuk dijadikan sebagai standar pengukuran untuk mengakurasikan hasil harga. Penelitian selanjutnya yang akan dilakukan adalah penambahan data kos yang lebih luas dan penerapan metode deep learning. Implementasikan lebih lanjut pada berbagai *platform*, salah satunya adalah dengan melakukan *deployment* ke dalam website juga akan dilakukan agar usulan dapat bermanfaat dan digunakan untuk keperluan yang lebih luas.

UCAPAN TERIMA KASIH

Ucapan terima kasih disampaikan kepada mata kuliah bahasa Indonesia yang diselenggarakan oleh UPN "Veteran" Jawa Timur sehingga artikel ini bisa ditulis dengan baik dan benar.

REFERENSI

[1] BPS. (2021). Jumlah Perguruan Tinggi, Mahasiswa, dan Tenaga Pendidik (Negeri dan Swasta) di Bawah Kementerian Riset, Teknologi dan Pendidikan Tinggi Menurut Kabupaten/Kota, 2019 dan 2020. [Online]. Diakses pada 14 Mei 2023 melalui https://jatim.bps.go.id/statictable/2021/09/06/2218/jumlah-perguruan-tinggi-mahasiswa-dan-tenaga-pendidik-negeri-dan-swasta-di-bawah-kementerian-riset-teknologi-dan-pendidikan-tinggi-menurut-kabupaten-kota-2019-dan-2020.html.

[2] H. Mukhlis, B. Ayshwarya, P. T. Nguyen, dan W. Hashim, "Boarding House Selection using SAW Method," July 2020.

[3] S. Suryana, "Sistem Rekomendasi Tempat Kos Mahasiswa Baru dengan Metode Naïve Bayes Berbasis Web," Jurnal Sistem Informasi Dan Komputer Terapan Indonesia (JSIKTI), vol. 3, no. 3, pp. 22-31, 2021. [Online]. Available: https://doi.org/10.33173/jsikti.107.

[4] A. B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, dan G. Oluwadara, "House Price Prediction using Random Forest Machine Learning Technique," Procedia Computer Science, vol. 199, pp. 806-813, 2021. [Online]. Tersedia: https://doi.org/10.1016/j.procs.2022.01.100.

[5] W. S. Pramana, M. Sudarma, dan I. N. S. Kumara, "Expert system and classical probability for setting up hotel's dynamic price level: A case of four-star hotel in Bali," International Journal of Electrical and Electronic Engineering and Telecommunications, vol. 9, no. 2, pp. 124-131, 2020. [Online]. Tersedia: https://doi.org/10.18178/IJEETC.9.2.124-131.

[6] W. T. Embaye, Y. A. Zereyesus, dan B. Chen, "Predicting the rental value of houses in household surveys in Tanzania, Uganda and Malawi: Evaluations of hedonic pricing and machine learning approaches," PLoS ONE, vol. 16, no. 2, pp. 1-20, February 2021. [Online]. Tersedia: https://doi.org/10.1371/journal.pone.0244953.

[7] M. K. Dahouda dan I. Joe, "A Deep-Learned Embedding Technique for Categorical Features Encoding," IEEE Access, vol. 9, pp. 114381-114391, 2021. [Online]. Tersedia: https://doi.org/10.1109/ACCESS.2021.3104357.

[8] Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends, 2(01), 20–28. https://doi.org/10.38094/jastt20165

[9] V. W. Siburian dan I. E. Mulyana, Prediksi Harga Ponsel Menggunakan Metode Random Forest, Prosiding Annual Research Seminar 2018, vol. 4, no. 1, hal. 144-147, 2018.

[10] Haifeng Wang and Dejin Hu, "Comparison of SVM and LS-SVM for Regression," 2005 International Conference on Neural Networks and Brain, Beijing, 2005, pp. 279-283, doi: 10.1109/ICNNB.2005.1614615.