

Penerapan *Random Forest Regression* Untuk Memprediksi Harga Jual Rumah Dan *Cosine Similarity* Untuk Rekomendasi Rumah Pada Provinsi Jawa Barat

Ega Sri Lestari¹, Ida Astuti*²

^{1,2}Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Gunadarma
Jl. Margonda Raya 100, Depok, Jawa Barat, 16424
¹egasrilestari@student.gunadarma.ac.id, ²astuti@staff.gunadarma.ac.id

*) Corresponding Author

(received: 01-07-22, revised: 11-08-22, accepted: 27-10-22)

Abstract

This study aims to implement a machine learning algorithm, namely *Random Forest Regression* in predicting house prices and the *Cosine Similarity* algorithm in providing house recommendations. The data is taken using web scraping technique, then the data will be processed using the CRISP-DM (*Cross-Industry Standard Process for Data Mining*) method with business understanding, data understanding, data preparation, modeling, evaluation and deployment stages. The results of the accuracy of the prediction process with tuning parameters are 85.29%, while the results of the recommendation accuracy are 89.99% on the test data. The research has succeeded in implementing a model in the form of a website that can be used by users in searching for house price needs in the province of West Java, then recommendations are given by switching the link to the rumah123.com website for more complete information. The website that has been built has been tested by inferential testing to get a precision value obtained by the system 75%, recall 100%, system accuracy 80%, while the *f-measure* is 86%, while the recommendation system gets a precision value obtained by the system 78%, recall 100%, 80% system accuracy, while the *f-measure* is 88%. User acceptance testing on the website gets a percentage of 89.29% with a very good category.

Keyword: *Random Forest Regression*., *Cosine Similarity*, *Prediction Recommendation*, *House prices*

Abstrak

Penelitian ini bertujuan untuk mengimplementasikan sebuah algoritma *machine learning* yaitu *Random Forest Regression* dalam memprediksi harga rumah dan algoritma *Cosine Similarity* dalam memberikan rekomendasi rumah. Data yang diambil menggunakan teknik *web scraping*, lalu data tersebut akan di olah menggunakan metode CRISP-DM (*Cross-Industry Standard Process for Data Mining*) dengan tahap *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation* dan *deployment*. Hasil akurasi dari proses prediksi dengan tuning parameter sebesar 85,29%, sedangkan hasil akurasi rekomendasi mendapatkan hasil 89,99% pada data uji. Penelitian berhasil mengimplementasikan model berupa website yang dapat digunakan oleh pengguna dalam mencari kebutuhan harga rumah di daerah provinsi Jawa Barat kemudian rekomendasi diberikan dengan peralihan link menuju website rumah123.com untuk informasi lebih lengkap. Website yang sudah dibangun telah diuji dengan pengujian inferensial mendapatkan nilai *precision* yang didapatkan oleh sistem 75%, *recall* 100%, akurasi sistem 80%, sedangkan *f-measure* 86%, sedangkan pada sistem rekomendasi mendapatkan nilai *precision* yang didapatkan oleh sistem 78%, *recall* 100%, akurasi sistem 80%, sedangkan *f-measure* 88%. Pengujian dengan *user acceptance test* pada website mendapatkan *persentase* sebesar 89.29% dengan kategori sangat baik.

Kata Kunci: *Random Forest Regression*, *Cosine Similarity*, *Prediksi*, *Rekomendasi*, *Harga*, *Rumah*

I. Pendahuluan

Provinsi Jawa Barat merupakan sebuah provinsi yang memiliki jumlah penduduk terbanyak dibandingkan dengan provinsi lain di Indonesia. Pertumbuhan penduduk ini mengakibatkan masyarakat harus memiliki tempat tinggal yang layak dengan harga yang terjangkau sesuai dengan kebutuhannya masing-masing.

Rumah merupakan suatu kebutuhan primer dari masyarakat dalam melindungi diri, tempat untuk beristirahat dan menyimpan beberapa barang. Seiring dengan perkembangan zaman, alat prediksi sebagai alat bantu telah dijadikan suatu bahan pertimbangan dalam mengambil keputusan pembelian [1]. Aplikasi prediksi dengan memberikan rekomendasi sangat cocok digunakan pada masyarakat untuk mendapatkan rumah yang diinginkan dengan cepat dan mudah.

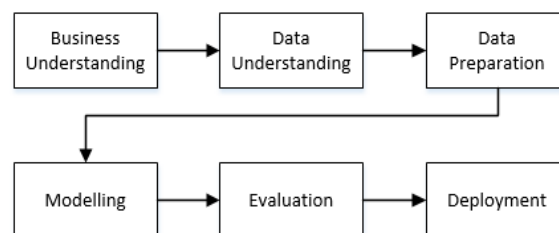
Secara sederhana *Machine Learning* ini adalah salah satu cabang ilmu kecerdasan buatan, khususnya dapat mempelajari tentang bagaimana komputer mampu belajar dari data untuk meningkatkan kecerdasannya [2]. Pada tahun 2020, terdapat penelitian dalam penggunaan metode *Moving Average* untuk memprediksi harga rumah di kota Bandung. Akurasi yang di hasilkan untuk pengujian ini 75% dengan nilai error MSE 15% [3]. Pada tahun 2021, telah dilakukan penelitian dengan menggunakan algoritma *Linear Regression* untuk memprediksi harga rumah. Hasil akurasi yang didapatkan sebesar 80% namun memiliki galat error yang cukup tinggi sebesar 259171.91[1].

Memberikan sistem rekomendasi kepada pengguna agar memudahkan dalam melakukan pencarian referensi rumah. Pada tahun 2020, Penelitian dalam menggunakan *Cosine Similarity* dalam mencari *keyword* pada sesuatu yang hilang menggunakan metode TF-IDF menghasilkan akurasi sebesar 88% [4].

Berdasarkan beberapa penelitian yang telah dilakukan sebelumnya, peneliti tertarik untuk melakukan penelitian prediksi harga rumah di provinsi Jawa Barat menggunakan algoritma *Random Forest Regression* serta memberikan rekomendasi referensi rumah menggunakan algoritma *Cosine Similarity*. Alasan menggunakan algoritma *Random Forest Regression* ini dikarenakan pada beberapa penelitian terkait algoritma ini memberikan hasil yang paling bagus untuk mengatasi permasalahan prediksi atau regresi lain.

II. Metodologi Penelitian

Metode penelitian ini akan menggunakan salah satu metodologi data science yaitu CRISP-DM (*Cross-Industry Standard Process for Data Mining*).



Gambar 1. Tahapan Penelitian

Berdasarkan Gambar 1 di atas penjabaran tahap penelitian yang akan dilakukan sebagai berikut:

Business Understanding

Masalah yang akan diselesaikan pada penelitian ini adalah mewujudkan hasil prediksi rumah berdasarkan input pengguna serta memberikan rekomendasi rumah yang berkaitan dengan profil pengguna tersebut. Metode yang cocok digunakan untuk menyelesaikan masalah ini menggunakan algoritma regresi yang dapat memprediksi nilai terhadap inputan yang diberikan serta menggunakan algoritma perhitungan jarak agar menghasilkan rekomendasi yang dapat disesuaikan dengan kemiripan data input.

Data Understanding

Pada tahap kedua akan melakukan proses untuk pengambilan data penelitian yaitu menggunakan teknik *web scrapping* pada website <https://www.rumah123.com/jual/jawa-barat/rumah/>. Tahap validasi ini diantaranya adalah mengganti nama kolom pada data, mengganti tipe data, membuang data yang tidak diperlukan. Proses selanjutnya adalah menelaah data yang bertujuan untuk menganalisa data secara statistika deskriptif agar dapat

melihat karakteristik data, kemudian melakukan pengecekan terhadap korelasi antar kolom untuk melihat keterhubungan kolom yang saling mempengaruhi. Tahap terakhir dalam *data understanding* adalah eksplorasi data dapat dilakukan dengan berbagai macam seperti membentuk pivot tabel, visualisasi diagram dan sebagainya.

Data Preparation

Melakukan tahap seleksi data untuk menentukan kolom yang memiliki pengaruh paling besar dengan target penelitian. Pembersihan data bertujuan untuk meminimalkan *noise*/data yang tidak lengkap sehingga data menjadi bersih dari bias/kesalahpahaman data. Pembersihan data ini dimulai dari pengecekan nilai kosong/*missing value*, lalu mengecek data duplikat dan yang terakhir adalah pengecekan *outlier*. Salah satu cara yang dilakukan untuk mengecek *outlier* pada data dengan metode *Interquartile Range* (IQR) dengan rumus sebagai berikut:

$$IQR = \text{Kuartil3} - \text{Kuartil1} \quad (1)$$

Lalu cek nilai *outlier* dengan syarat data < Kuartil 1 - 1.5 * IQR atau data > Kuartil 3 + 1.5 * IQR. Konstruksi data bertujuan untuk menambahkan fitur, normalisasi data dan melakukan transformasi data non-numerik menjadi numerik. Pada proses normalisasi akan menggunakan metode *MinMaxScaler* dengan rumus sebagai berikut:

$$\text{newdata} = \frac{(\text{data}-\text{min}) \times (\text{newmax}-\text{newmin})}{(\text{max}-\text{min})} + \text{newmin} \quad (2)$$

Keterangan:

Newdata = Data baru hasil normalisasi.

Data = Data yang akan dinormalisasi.

Min = Nilai terkecil dari data yang akan dinormalisasi.

Max = Nilai terbesar dari data yang akan dinormalisasi.

Newmin = Batas nilai terkecil normalisasi.

Newmax = Batas nilai terbesar normalisasi.

Modeling

Pemodelan data latih dengan algoritma *Random Forest Regression* untuk memprediksi harga rumah dan *Cosine Similarity* untuk rekomendasi rumah berdasarkan hasil prediksi. Pemisahan data yang akan dilakukan menggunakan 3 skenario diantaranya skenario pertama 80% data latih dan 20% uji, skenario kedua 75% data latih dan 25% data uji, skenario ketiga 70% data latih dan 30% data uji.

Random Forest merupakan salah satu metode yang digunakan untuk klasifikasi dengan membangun banyak pohon klasifikasi. Metode ini dapat meningkatkan hasil akurasi, dengan cara membangkitkan simpul anak untuk setiap *node* (simpul di atasnya) dan dilakukan pemilihan secara acak [5]. Pembentukan akar pohon akan diambil berdasarkan nilai terkecil *splitting criteria* pada masing-masing fitur yang digunakan. Pada masalah regresi menggunakan *Random Forest* disarankan untuk menggunakan perhitungan MSE dalam menentukan *splitting criteria* pohon dengan menggunakan rumus sebagai berikut:

$$MSE_n = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y}_n)^2 \quad (3)$$

Keterangan:

MSE_n = Nilai MSE pada pohon ke-n.

N = Jumlah sampel pada pohon ke-n.

$Y_i Y_i$ = Nilai sampel ke-I pada pohon ke-n.

\bar{Y}_n = Nilai rata-rata sampel pohon ke-n.

Prediksi pada kasus regresi diambil dari nilai rata-rata dari setiap pohon Rumus untuk menghitung nilai rata-rata seluruh prediksi pohon sebagai berikut:

$$\hat{Y}_i = \frac{1}{N_{tree}} \sum_{n=1}^{N_{tree}} \hat{Y}_n \quad (4)$$

Keterangan:

\hat{Y}_i = Hasil prediksi akhir.

N_{tree} = Total jumlah pohon pada *Random Forest*.

\hat{Y}_n = Hasil prediksi pohon ke-n.

Pengujian *performance* dari model yang telah dibentuk akan dibandingkan dengan menggunakan tuning parameter dimana terdapat 2 model yang akan dibandingkan yaitu model dengan parameter *default* dan model dengan parameter pilihan dari tuning. Metode tuning parameter yang akan digunakan adalah *Grid Search CV* untuk data pada setiap *fold* kemudian akan digunakan untuk membuat model baru yang lebih baik daripada model sebelum dilakukan pengoptimalan parameter [6].

Cosine Similarity ini akan mengukur sudut antara perkalian dua vektor yang diproyeksikan dalam ruangan multidimensi. Nilai *Cosine Similarity* yang paling tinggi akan menjadi hasil rekomendasi kepada *user* [7]. Rumus yang akan digunakan untuk perhitungan *Cosine Similarity* sebagai berikut:

$$\text{sim}(Q, D_i) = \frac{Q \cdot D}{|Q| |D|} = \frac{\sum_{y=1}^n (w_{Q,y} \times w_{i,y})}{\sqrt{\sum_{y=1}^n w_{Q,y}^2 \times \sum_{y=1}^n w_{i,y}^2}} \quad (5)$$

Keterangan:

Q = Vektor Q, dibandingkan kemiripan datanya.

D = Vektor D, dibandingkan kemiripan datanya.

$Q \cdot D$ = Dot *product* antara vektor Q dan D.

$|Q|$ = Panjang vektor Q.

$|D|$ = Panjang vektor D.

$|Q| |D|$ = *Cross product* antara vektor Q dan vektor D.

$w_{Q,y}$ = Term ke-y yang terdapat pada data W_Q .

$w_{i,y}$ = Term ke-y yang terdapat pada dokumen W_i .

Evaluation

Tahap evaluasi akan melakukan pengukuran akurasi data terhadap hasil prediksi dan rekomendasi yang diberikan menggunakan data testing pada masing-masing model. Sebuah model *Machine Learning* dikatakan baik harus memberikan hasil output yang akurat dalam rangka untuk pengambilan keputusan berdasarkan hasil yang diberikan [8]. Pengukuran yang akan dilakukan menggunakan metrik MSE (*Mean Squared Error*), MAE (*Mean Absolute Error*) dan *R-Squared*.

MSE dapat dikatakan sebagai rata-rata kuadrat kesalahan yang dapat dihitung dengan menjumlahkan semua kesalahan atau error hasil prediksi atau peramalan pada setiap data input yang dimasukkan kemudian dikuadratkan dan membaginya dengan jumlah periode peramalan [9]. Rumus untuk menghitung MSE sebagai berikut:

$$MSE = \frac{1}{n} \sum_{i=1}^n (X_i - F_i)^2 \quad (6)$$

Keterangan:

X_i = Data aktual pada periode ke-i.

F_i = Nilai hasil ramalan atau prediksi pada periode ke-i.

n = Banyaknya data sampel.

Pada pengukuran model MAE setiap selisih error akan diambil nilai mutlak nya untuk selanjutnya di jumlahkan.

Rumus untuk menghitung MAE sebagai berikut:

$$MAE = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j| \quad (7)$$

Keterangan:

n = Banyaknya data sampel.

y_j = Nilai aktual untuk sample j.

\hat{y}_j = Nilai prediksi untuk sample j.

Pengukuran model dengan *R-squared* adalah proporsi variasi dalam variabel dependen yang dapat dijelaskan oleh variabel-variabel independen nya [9]. Rumus *R-Squared* sebagai berikut:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(Y_{aktual} - Y_{pred})^2}{\sum(Y_{aktual} - \bar{Y})^2} \quad (8)$$

Keterangan:

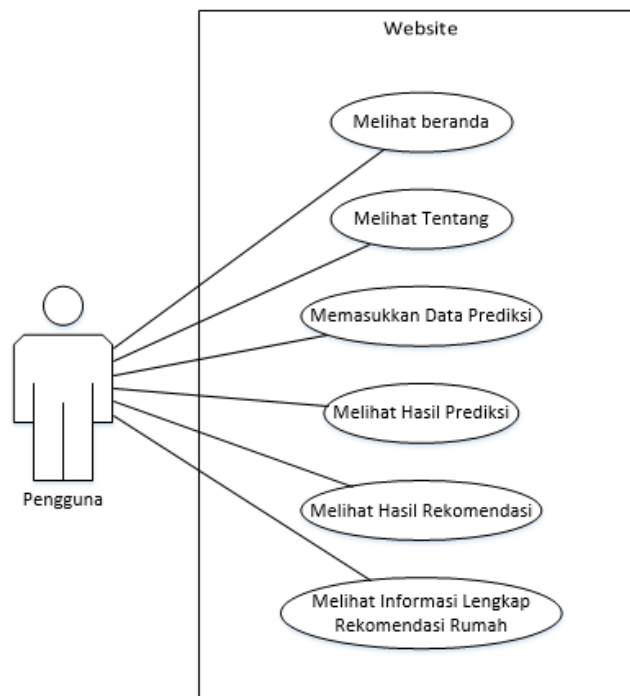
Y_{aktual} = Data uji.

Y_{pred} = Data uji hasil prediksi.

\bar{Y} = Rata-rata data uji.

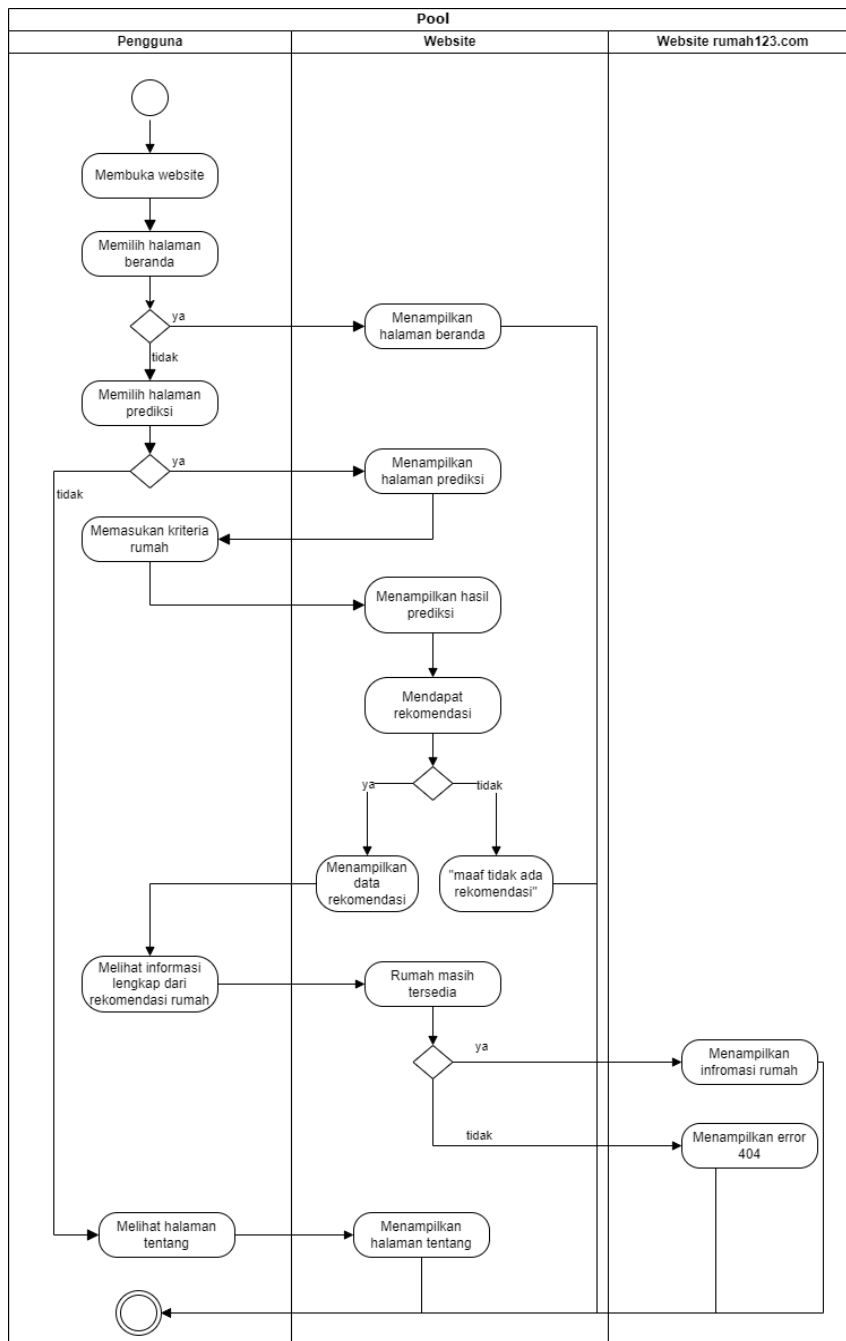
Deployment

Tahap *deployment* dimana akan mengimplementasikan hasil model yang telah disimpan dan hasil rekomendasi ke dalam website menggunakan *framework* yang disediakan *Python* yaitu *flask*. Tahap terakhir pada penelitian adalah pengujian website prediksi harga rumah yang akan menggunakan pengujian inferensial dan pengujian *user acceptance*.



Gambar 2. Use Case Website

Berdasarkan Gambar 2 di atas merupakan *use case* dari website yang akan dibangun, *user* yang dapat menggunakan website ini hanya pengguna. Kegiatan yang dapat dilakukan oleh pengguna website meliputi melihat beranda yang menampilkan informasi website dan tipe-tipe rumah, kemudian dapat melihat halaman tentang yang menampilkan deskripsi singkat pembuat website. Pengguna juga dapat memasukkan data prediksi yang akan menghasilkan nilai prediksi dan rekomendasi dari implementasi model. Kegiatan lain yang dapat dilakukan pengguna adalah melihat hasil prediksi dan hasil rekomendasi.



Gambar 3. Activity Diagram Website

Berdasarkan Gambar 3 di atas aktifitas pertama yang akan dilakukan oleh pengguna adalah membuka halaman beranda website prediksi, lalu website akan menampilkan halaman beranda. Apabila pengguna tidak membuka halaman beranda dan membuka halaman prediksi maka website akan menampilkan halaman prediksi, lalu pengguna diharuskan untuk memasukkan kriteria rumah sehingga website akan menampilkan hasil prediksi. Jika data yang dimasukkan pengguna mendapatkan hasil rekomendasi maka website akan menampilkan data rekomendasi, namun apabila data yang dimasukkan tidak mendapatkan rekomendasi maka tidak akan menampilkan hasil rekomendasi. Apabila pengguna sudah melihat hasil prediksi dan rekomendasi maka pengguna juga dapat melihat informasi lengkap mengenai rumah rekomendasi yang diberikan jika rumah masih tersedia maka website rumah123.com akan menampilkan informasi lebih lengkap terkait rumah tersebut seperti detail dari lokasi rumah, gambar dari rumah yang ditawarkan dan informasi lainnya mengenai rumah yang ditawarkan,

apabila rumah sudah tidak tersedia maka website rumah123.com akan menampilkan halaman *error* 404 dikarenakan ada kemungkinan bahwa rumah sudah tidak di iklankan pada website rumah123.com.

III. Hasil dan Pembahasan

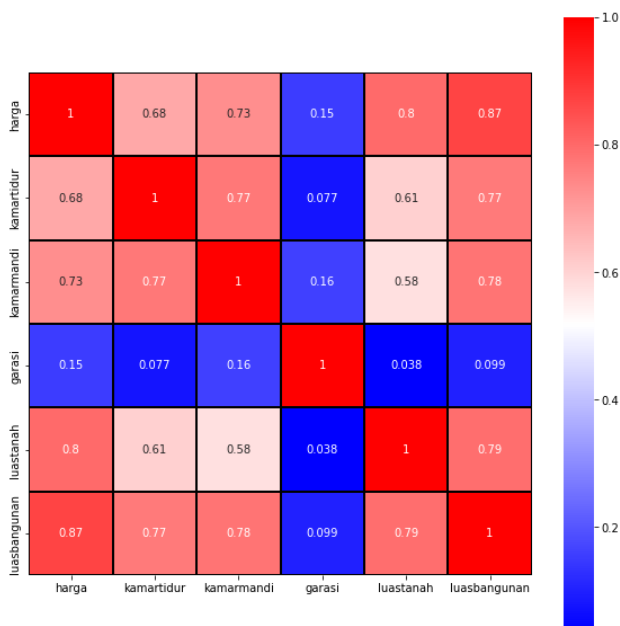
Hasil pengumpulan data dari proses *web scrapping* pada website rumah123.com mendapatkan data sebanyak 1004 baris dan 9 kolom yang diambil dari tanggal 1 Maret 2022 hingga 5 Maret 2022 hanya wilayah provinsi Jawa Barat.

Tabel 1. Hasil Pengumpulan Data

no.	nama	daerah	harga	kamar tidur	kamar mandi	garasi	luas tanah	luas bangunan	link
1	Cluster Sawangan.	Depok	485 Juta	2	1	1	60	45	/properti/depok/hos9945070/
2	Rumah Mewah Depok	Depok	1,42 Miliar	4	3	2	93	130	/properti/depok/hos9612079/
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1003	New Diamond	Depok	800 Juta	2	2	1	97	45	/properti/depok/hos9748561/
1004	Hunian Cantik Depok	Depok	950 Juta	3	3	2	95	100	/properti/depok/hos9943216/

Berdasarkan Tabel 1 dapat dilihat bahwa pada kolom harga masih menggunakan tipe data string sehingga harus diubah ke dalam bentuk integer menggunakan metode *Regex* dengan syarat kata “Juta” akan dikalikan dengan 1E6 sedangkan kata “Miliar” akan dikalikan dengan 1E9. Pada proses validasi ini akan melakukan penghapusan terhadap data yang tidak dibutuhkan seperti mengecek data dengan nama rumah yang mempunyai kata “kos”, “kost”, “koston” data tersebut akan dihapus karena akan memberikan efek ketidakstabilan pada data. Lakukan penghapusan terhadap data yang tidak diperlukan kemudian dilakukan cek terhadap jumlah data yang sudah bersih dari penyimpangan sehingga jumlah data sekarang sebanyak 994 baris data dengan 9 kolom.

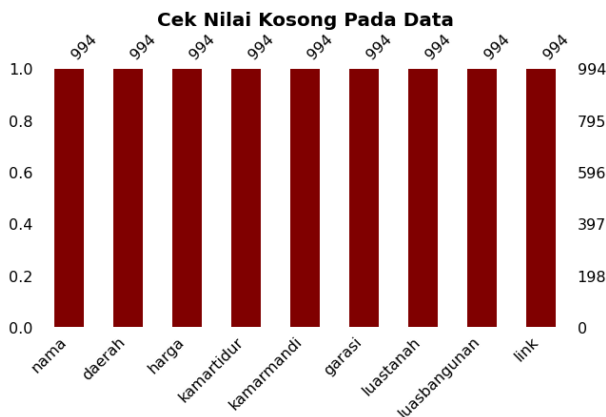
Korelasi Data Prediksi Harga Rumah Provinsi Jawa Barat



Gambar 4. Hasil Korelasi Data

Berdasarkan Gambar 4 di atas terlihat bahwa harga rumah memiliki korelasi paling tinggi dengan luas bangunan yaitu 0.87 yang berarti jika harga rumah mengalami peningkatan maka hal tersebut sangat dipengaruhi oleh luas bangunan. Pada korelasi antara harga rumah dengan jumlah garasi adalah korelasi yang kecil akan tetapi memiliki nilai yang positif berarti harga rumah sulit untuk di pengaruhi oleh jumlah garasi dari rumah tersebut. Proses

seleksi data menghasilkan hasil yang sama dengan korelasi sehingga menyatakan bahwa fitur kolom luas bangunan memiliki kontribusi yang paling berpengaruh dalam memprediksi harga rumah.



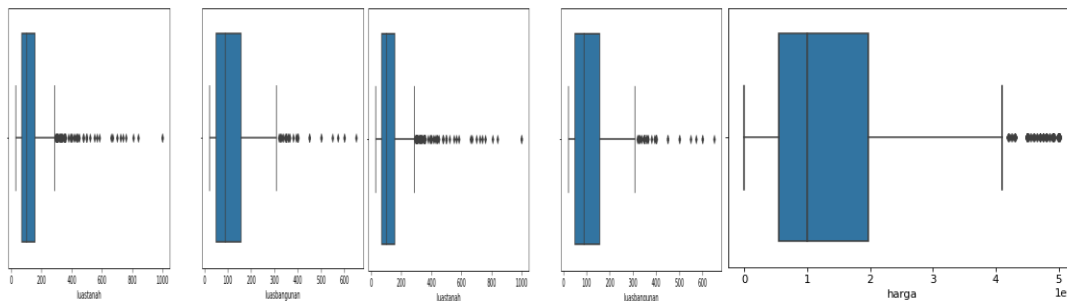
Gambar 5. Cek Nilai Hilang

Berdasarkan Gambar 5 di atas pada proses pembersihan data akan melakukan perbaikan, penghapusan dan pengabaian dari *noise* yang ada pada data. Semua jumlah nilai pada setiap atribut memiliki jumlah yang sama yaitu masing-masing berjumlah 994 data sehingga tidak dibutuhkan untuk melakukan penanganan terhadap nilai hilang. Jika pada data terdapat nilai yang hilang maka harus segera ditangani sesuai dengan karakteristik data tersebut. Pada setiap penanganan nilai hilang memiliki kriteria yang dapat disesuaikan dengan data.

```
Ukuran data dengan duplikat :  
Kolom: 9  
Baris: 994  
Ukuran data tanpa duplikat :  
Kolom: 9  
Baris: 990
```

Gambar 6. Cek Data Duplikat

Berdasarkan Gambar 6 di atas jika sudah selesai mengecek nilai kosong pada data maka lakukan pengecekan terhadap data duplikat, dari hasil yang telah dilakukan terdapat 4 data duplikat sehingga data duplikat ini dapat dihapus sehingga jumlah data tanpa duplikat sebanyak 990 baris dengan 9 kolom. Metode yang akan digunakan untuk mengecek nilai *outlier* adalah IQR dan *boxplot*.



Gambar 7. Outlier Pada Seluruh Data

Berdasarkan Gambar 7 di atas kolom yang mempunyai *outlier* dengan jumlah terbanyak dari seluruh data yang digunakan. Jika di hitung menggunakan rumus IQR pada kolom luas tanah memiliki *outlier* paling banyak yaitu sebanyak 85 data, data yang memiliki *outlier* akan dihapus barisnya sehingga data yang sudah bersih siap untuk di modeling memiliki 848 baris dan 9 kolom.

Proses selanjutnya adalah melakukan konstruksi data dengan transformasi kolom daerah menjadi tipe data integer. Transformasi ini akan menggunakan *function* LabelEncoder pada *Python*. Proses transformasi ini sangat penting dilakukan karena *Machine Learning* tidak dapat mengolah data dengan sifat *string* atau *object*.


```
[ 'Bandung' 'Banjar' 'Bekasi' 'Bogor' 'Cianjur' 'Cikarang' 'Cimahi'
  'Cirebon' 'Depok' 'Indramayu' 'Karawang' 'Kuningan' 'Purwakarta' 'Subang'
  'Sumedang' ]
[ 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 ]
```

Gambar 8. Hasil Pelabelan Data

Berdasarkan Gambar 8 di atas dapat dilihat bahwa sekarang data kota sudah berubah menjadi angka yang diurutkan berdasarkan abjad daerah tersebut. Proses normalisasi menggunakan MinMaxScaler untuk mengubah rentang pada data sehingga semua kolom pada saat modeling akan memiliki rentang yang sama yaitu 0-1.

daerah	harga	kamartidur	kamarmandi	garasi	luastanah	luasbangunan
0	0.571429	0.199544	0.666667	0.2	0.5	0.166667
1	0.571429	0.200051	0.666667	0.2	0.5	0.166667
2	0.571429	0.199291	0.666667	0.2	0.5	0.166667
3	0.571429	0.219562	0.666667	0.2	0.5	0.166667
4	0.571429	0.173952	0.666667	0.2	0.5	0.135659
...
843	0.571429	0.181553	0.333333	0.0	0.5	0.236434
844	0.571429	0.201824	0.666667	0.2	0.5	0.178295
845	0.000000	0.174712	0.666667	0.2	1.0	0.124031
846	0.000000	0.097681	0.666667	0.2	0.5	0.116279
847	0.142857	0.379197	0.666667	0.2	0.0	0.383721

Gambar 9. Hasil Normalisasi Data

Berdasarkan Gambar 9 di atas hasil data yang telah dinormalisasi sudah memiliki rentang yang sama antar kolom. Jika sudah melakukan proses data konstruksi maka data sudah siap untuk dimasukkan ke dalam tahap model dengan algoritma *Random Forest Regression*.

Tabel 2. Hasil Pemisahan Data

Kriteria	Data Latih	Data Uji
80:20	659	165
75:25	618	206
70:30	577	247

Berdasarkan Tabel 2 di atas masing-masing skenario akan dilakukan pelatihan dengan algoritma *Random Forest Regression* dengan parameter yang sama. Pada tahap modeling algoritma *Random Forest Regression* hanya menggunakan parameter `n_estimators`, sedangkan untuk parameter lain di gunakan secara *default*.

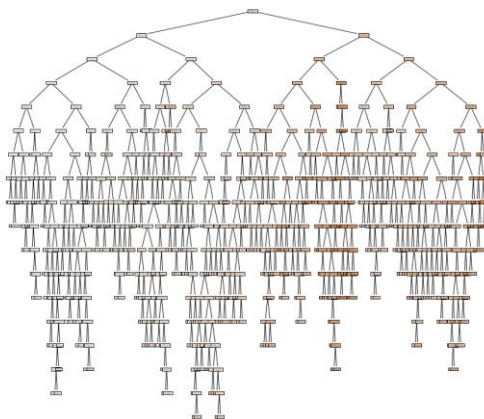
```
forest80 = RandomForestRegressor(n_estimators=300, random_state=42)
forest80.fit(X_train80, y_train80)
forest75 = RandomForestRegressor(n_estimators=300, random_state=42)
forest75.fit(X_train75, y_train75)
forest70 = RandomForestRegressor(n_estimators=300, random_state=42)
forest70.fit(X_train70, y_train70)
```

Pada code di atas merupakan proses melakukan pelatihan terhadap data training dengan algoritma *Random Forest Regression* menggunakan *Python*. Parameter yang ditentukan hanya jumlah pohon yang terbentuk yaitu 300 sedangkan parameter lain seperti `bootstrap`, `max_features` dan lainnya ditentukan secara *default*. Pada setiap skenario akan dilakukan pelatihan dengan algoritma *Random Forest Regression* dengan parameter yang sama. Pada tuning parameter ini akan memanfaatkan library *GridSearchCV* dalam memberikan parameter terbaik.

Tabel 3. Akurasi Model 3 Skenario

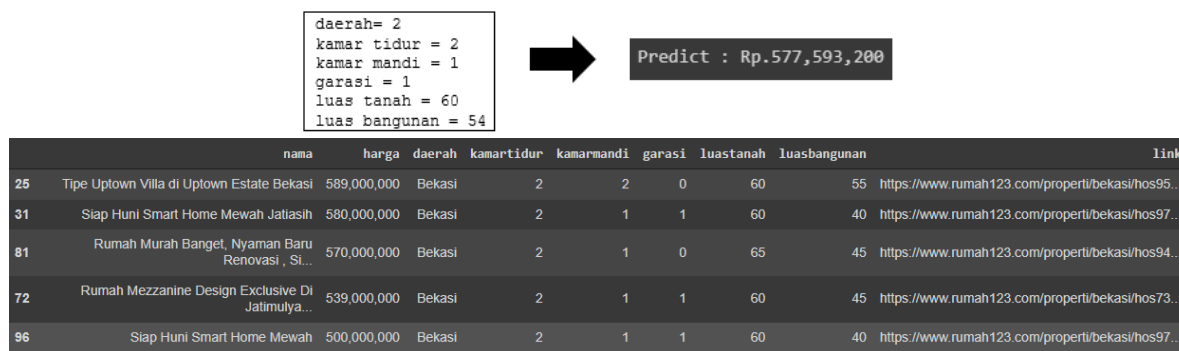
Skenario	Sebelum Tuning	Setelah Tuning
80:20	78.93	80.92
75:25	80.30	85.29
70:30	79.17	83.20

Berdasarkan Tabel 3 di atas akurasi data uji tertinggi didapatkan oleh kriteria 75% data latih dan 25% data uji dengan hasil prediksi menggunakan parameter metode *Grid Search CV* naik sekitar 4.99%.



Gambar 10. Hasil Algoritma *Random Forest Regression* Setelah Tuning Parameter

Berdasarkan Gambar 10 di atas merupakan hasil pohon yang berhasil dibentuk oleh model dengan split data 75% data latih dan 25% data uji setelah menggunakan parameter *gridsearchCV*. Pada proses pemodelan algoritma *Cosine Similarity* akan menggunakan data split dengan kriteria 75% data latih dan 25% data uji. Pembentukan data untuk memberikan hasil rekomendasi berdasarkan nilai *Cosine Similarity* antara data latih dengan data input yang dimasukkan pengguna. Proses *content based filtering* sebagai sebuah metode tambahan untuk pelengkap dari hasil rekomendasi rumah yang diberikan. Metode *filtering* tambahan ini akan memberikan efek untuk memperkecil lingkup rekomendasi sehingga hasil rekomendasi yang diberikan akan semakin akurat mendekati profil pengguna. Fitur yang akan di gunakan adalah 3 fitur yang memiliki kontribusi penting bagi pengguna dalam memilih harga rumah yaitu daerah, harga dan luas tanah.



Gambar 11. Proses *Cosine Similarity*

Berdasarkan Gambar 11 di atas merupakan proses perolehan data prediksi dan rekomendasi yang akan diberikan. Dari data input pengguna maka akan diberikan hasil harga prediksi rumah berdasarkan kriteria tersebut. Harga prediksi ini akan di lakukan proses dengan algoritma *Cosine Similarity* dan menggunakan filtering 3 faktor tambahan yaitu harga, kota dan luas tanah. hasil yang diberikan oleh perhitungan *Cosine Similarity* ini cukup dekat dengan data input yang dimasukkan oleh pengguna. Terlihat dari kota yang dipilih, luas bangunan yang

hampir mendekati data input, dan sebagainya. Proses terakhir mengecek akurasi dari model yang sudah dibentuk dengan pemisahan 75% data latih dan 25% data uji pada model prediksi dan model rekomendasi. Metrik yang akan digunakan adalah MSE, MAE, dan *R-Squared* menggunakan bahasa pemrograman *Python*.



Gambar 12. Hasil Evaluasi Model

Berdasarkan Gambar 12 di atas dengan menggunakan *Python* pada seluruh dataset yang digunakan untuk pelatihan *random forest* setelah *tuning parameter* mengalami penurunan error baik pada metrik MSE maupun MAE. Pada sistem rekomendasi dengan menggunakan algoritma *Cosine Similarity* mendapatkan hasil 6.26 pada MSE dan 2.05 pada MAE sehingga dapat dikatakan bahwa rekomendasi yang dihasilkan cukup baik. Hasil evaluasi pada algoritma *Random Forest Regression* mengalami peningkatan yang cukup baik setelah melakukan proses *tuning parameter* dengan hasil sebesar 85.29% dan pada algoritma *Cosine Similarity* sebesar 89.99%.

Proses terakhir pada penelitian ini adalah *deployment* yang akan melakukan implementasi model yang sudah disimpan menjadi sebuah sistem berbasis website agar dapat diakses dan digunakan oleh masyarakat dalam membantu menemukan rumah yang diinginkan. Pada proses pembuatan website sebagai implementasi dari hasil analisis ini akan menggunakan *framework Flask* yang ada pada bahasa pemrograman *Python* sebagai *backend*, lalu menggunakan HTML dan CSS sebagai *frontend* dari website yang akan dibangun.



Gambar 13. Tampilan Halaman Beranda

Berdasarkan Gambar 13 merupakan tampilan halaman beranda pada website prediksi. Pada halaman ini akan menampilkan informasi tentang tipe-tipe rumah mulai dari rumah 1 lantai hingga 5 lantai.

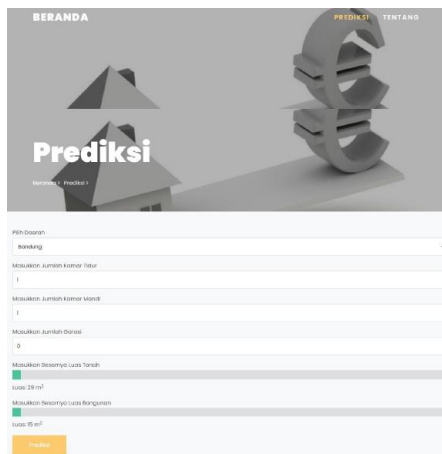


Ega Sri Lestari, Hobi Menulis

Saya mahasiswa semester 8 pada Universitas Guadalupe Jurusan Sistem Informasi. Saya sangat senang untuk mengembangkan diri saya pada bidang data science. Saya sangat senang yang menggunakan media, digital, di dapat berbagai macam ilmu mengenai data science, dan bertanggung jawab. Saya mempunyai hobi di baca tentang programming yaitu pemrograman.

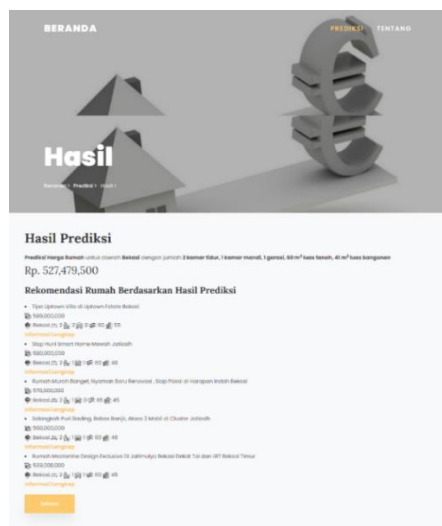
Gambar 14. Tampilan Halaman Tentang

Berdasarkan Gambar 14 pada tampilan halaman tentang akan menampilkan informasi singkat mengenai deskripsi penulis dan deskripsi penggunaan website prediksi untuk pada pengguna aplikasi.



Gambar 15. Tampilan Halaman Prediksi

Berdasarkan Gambar 15 merupakan hasil tampilan halaman prediksi yang bertujuan untuk memberikan input kriteria rumah yang akan diprediksi.



Gambar 16. Tampilan Halaman Hasil

Berdasarkan Gambar 16 di atas merupakan tampilan hasil prediksi harga rumah dan halaman ini akan ditampilkan informasi kriteria yang dipilih beserta hasil prediksi harga rumah. Setelah menyelesaikan hosting maka website sudah dapat diakses secara online oleh masyarakat dengan alamat url: <https://websiteprediksinhargarumahjabar.herokuapp.com/>.

Pengujian inferensial akan melihat kinerja model yang diberikan setelah dilakukan *deployment*. Terdapat 10 data uji yang akan digunakan untuk menguji kinerja prediksi model dan rekomendasi model yang diberikan.

Tabel 4. Data Pengujian Inferensial Model Prediksi

data uji ke-	harga	prediksi	selisih	label prediksi	label aktual
data uji ke-1	365000000	355726000	9274000	relevan	relevan
data uji ke-2	550000000	484598200	65401800	tidak relevan	tidak relevan
data uji ke-3	567000000	591194300	24194300	relevan	relevan
data uji ke-4	380000000	352048500	27951500	relevan	relevan
data uji ke-5	314000000	320622500	6622500	relevan	tidak relevan
data uji ke-6	1200000000	923332500	276667500	tidak relevan	tidak relevan
data uji ke-7	489000000	450581500	38418500	relevan	relevan
data uji ke-8	412000000	448487500	36487500	relevan	relevan
data uji ke-9	457000000	482225500	25225500	relevan	relevan
data uji ke-10	350000000	303722500	46277500	relevan	tidak relevan

Berdasarkan Tabel 4 di atas pada kolom selisih di dapatkan dari hasil nilai *absolute* pengurangan dari harga asli dengan harga prediksi yang diberikan. Dari kolom selisih tersebut maka akan dihitung nilai rata-rata selisih harga yang diberikan antara harga asli dengan harga prediksi.

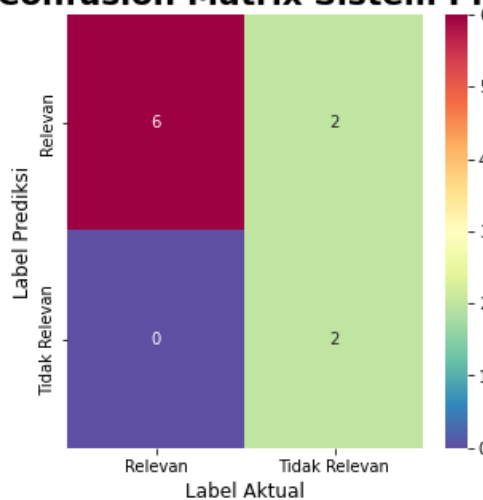
Pada label prediksi diberi sebuah kondisi:

- Relevan: apabila nilai selisih \leq rata-rata selisih.
- Tidak relevan: apabila nilai selisih $>$ rata-rata selisih.

Pada label aktual diberi sebuah kondisi berdasarkan beberapa asumsi pengguna yang menggunakan sistem:

- Relevan: apabila nilai selisih $<$ 40.000.000.
- Tidak relevan: apabila nilai selisih $>$ 40.000.000.

Hasil Confusion Matrix Sistem Prediksi



Gambar 17. Confusion Matrix Sistem Prediksi

Berdasarkan Gambar 17 diatas merupakan tabel tipe kesalahan dari *confusion matrix* yang sudah terbentuk. Beberapa pengukuran yang akan di lakukan berdasarkan hasil tabel di atas diantaranya *precision*, *recall*, akurasi, dan *F-measure*.

$$precision = \frac{TP}{TP+FP} = \frac{6}{6+2} = 0.75$$

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} = \frac{6+2}{6+2+2+0} = 0.8$$

$$recall = \frac{TP}{TP+FN} = \frac{6}{6+0} = 1.00$$

$$Fmeasure = 2 \times \frac{precision \times recall}{precision+recall} = \frac{2 \times 0.75 \times 1}{0.75+1} = 0.86$$

Tabel 5. Data Pengujian Inferensial Model Rekomendasi

data uji ke-	total rekomendasi	label rekomendasi	label aktual
data uji ke-1	5	relevan	relevan
data uji ke-2	5	relevan	tidak relevan
data uji ke-3	5	relevan	tidak relevan
data uji ke-4	5	relevan	relevan
data uji ke-5	5	relevan	relevan
data uji ke-6	1	relevan	relevan
data uji ke-7	4	relevan	relevan
data uji ke-8	5	relevan	relevan
data uji ke-9	4	relevan	relevan
data uji ke-10	0	tidak relevan	tidak relevan

Berdasarkan Tabel 5 merupakan data label yang diberikan sistem kepada pengguna. Total rekomendasi yang diberikan oleh sistem maksimal adalah 5 data rekomendasi.

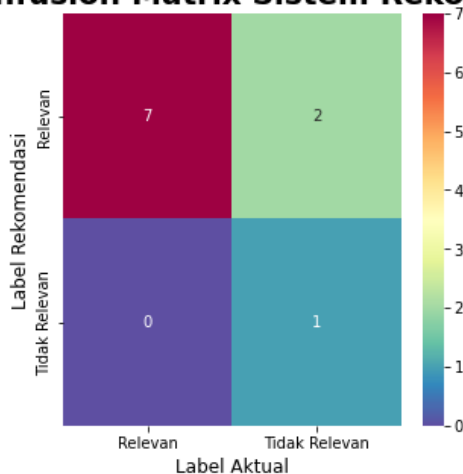
Pada label rekomendasi diberikan sebuah kondisi:

- Label relevan: apabila rekomendasi yang diberikan sistem terhadap profil pengguna jika terdapat minimal 1 data rekomendasi.
- Label tidak relevan: apabila tidak ada data rekomendasi yang diberikan oleh sistem.

Pada label aktual diberikan sebuah kondisi sebagai berikut:

- Label relevan: jika data rekomendasi yang diberikan setidaknya memiliki 4 kriteria yang sesuai dengan profil pengguna.
- Label tidak relevan: apabila hasil rekomendasi yang diberikan tidak sesuai dengan profil pengguna.

Hasil Confusion Matrix Sistem Rekomendasi



Gambar 18. *Confusion Matrix* Sistem Rekomendasi

Berdasarkan Gambar 18 diatas merupakan metrik tipe kesalahan confusion matrix yang digunakan pada sistem rekomendasi. Beberapa pengukuran yang akan di lakukan berdasarkan hasil tabel di atas diantaranya precision, recall, akurasi, dan F-measure [10].

$$precision = \frac{TP}{TP+FP} = \frac{7}{7} = 0.78$$

$$recall = \frac{TP}{TP+FN} = \frac{7}{7+0} = 1.00$$

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} = \frac{7+1}{7+2+1+0} = 0.8$$

$$Fmeasure = 2 \times \frac{precision \times recall}{precision+recall} = \frac{2 \times 0.78 \times 1}{0.78+1} = 0.88$$

Pengujian *User acceptance* dilakukan untuk mengetahui tanggapan dari pengguna pada website yang sudah dibuat. Pengisian kuesioner ini dilakukan dengan 20 responden untuk mengukur kepuasan pengguna terhadap hasil prediksi dan rekomendasi website.

Tabel 6. Hasil Pengujian *User acceptance*

Pertanyaan	Penilaian					Skor					Jumlah
	SS	S	N	TS	STS	SS	S	N	TS	STS	
Apakah menu-menu pada website mudah di pahami?	12	6	2	0	0	60	24	6	0	0	90
Apakah performa website baik dalam memberikan hasil harga prediksi?	11	5	4	0	0	55	20	12	0	0	87
Apakah website dapat memberikan hasil rekomendasi yang akurat?	10	7	3	0	0	50	28	9	0	0	87
Apakah fitur-fitur inputan seperti pemilihan daerah, luas tanah dan lain-lain pada website sudah baik?	15	5	0	0	0	75	20	0	0	0	95
Apakah website sudah sesuai dengan kebutuhan untuk melakukan prediksi dan memberikan rekomendasi harga rumah?	11	7	2	0	0	55	28	6	0	0	89
Apakah hasil rekomendasi yang ditampilkan oleh website sudah sesuai dengan keinginan anda?	11	8	1	0	0	55	32	3	0	0	90
Apakah link informasi selengkapnya pada website berjalan dengan baik?	10	7	3	0	0	50	28	9	0	0	87

Berdasarkan Tabel 6 di atas kolom skor didapatkan dari hasil perkalian penilaian dengan bobot yang di tentukan sesuai dengan label. Kolom jumlah didapatkan dari hasil penjumlahan skor observasi hasil kuesioner yang sudah di beri bobot masing-masing label.

$$Y = \frac{Skor\ yang\ diobservasi}{Skor\ yang\ diharapkan} \times 100\%$$

Skor yang diharapkan pada pengujian ini adalah $7 \times 100 = 700$ dan $7007 \times 100 = 700$, nilai 100 didapatkan karena pada pengujian bobot tertinggi adalah label SS dengan nilai 5 kemudian di kalikan dengan 20 responden jadi 100 adalah nilai tertinggi.

$$Y = \frac{90+87+87+95+89+90+87}{700} \times 100\% = \frac{625}{700} \times 100\% = 89,29\%$$

Berdasarkan hasil yang diperoleh dapat dilihat bahwa 20 pengguna website prediksi menilai bahwa hasil yang diberikan oleh aplikasi cukup mudah digunakan dan bermanfaat untuk membantu pengguna dalam menemukan rekomendasi dan prediksi harga rumah yang diinginkan.

Semua pengujian yang sudah di lakukan menyatakan bahwa model yang digunakan pada penelitian ini sudah baik karena sudah di atas nilai *cut off* (50%) yang pada umumnya digunakan, model dikatakan tidak baik apabila menghasilkan akurasi dibawah nilai *cut off* tersebut [11]. Hal ini berarti bahwa model sudah mampu dalam memberikan hasil prediksi dan rekomendasi yang akurat sebesar 89,29%.

IV. Kesimpulan

Website Aplikasi prediksi harga rumah dan rekomendasi rumah telah berhasil dibuat, untuk prediksi menggunakan algoritma *Random Forest Regression* sedangkan proses rekomendasi menggunakan algoritma *Cosine Similarity* dengan *content based filtering* pada fitur daerah, luas tanah, dan harga rumah. Metode yang digunakan adalah CRISP-DM (*Cross-Industry Standard Process for Data Mining*) dengan tahap *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation* dan *deployment*. Proses prediksi dilakukan menggunakan algoritma *Random Forest Regression* dengan akurasi tertinggi terdapat pada penggunaan 75% data latih dan 25% data uji. Berdasarkan hasil evaluasi model yang telah dilakukan pelatihan mendapatkan hasil akurasi prediksi sebesar 85,29%, sedangkan pada evaluasi rekomendasi mendapatkan hasil 89,99%.

Pada hasil pengujian inferensial website, sistem prediksi memperoleh nilai *precision* sebesar 75%, *recall* 100%, dan akurasi 80%, sehingga dapat dikatakan bahwa model prediksi telah bekerja dengan baik dalam memberikan hasil prediksi harga rumah dengan data uji yang diberikan. Pada sistem rekomendasi rumah dengan uji inferensial mendapatkan nilai *precision* sebesar 78%, *recall* 100% dan akurasi sebesar 80%, sehingga dapat dikatakan bahwa model rekomendasi telah bekerja dengan baik dalam memberikan hasil rekomendasi rumah dengan data uji yang diberikan. Pengujian *user acceptance* website mendapatkan *persentase* sebesar 89,29% disimpulkan bahwa pengguna memperoleh kemudahan dalam menggunakan website dan dapat bermanfaat untuk mengetahui prediksi harga rumah dan rekomendasi rumah di Provinsi Jawa Barat.

Daftar Pustaka

- [1] Andi Saiful, Septi Andryana, Aris Gunaryati, "Prediksi Harga Rumah Menggunakan Web Scrapping Dan Machine Learning dengan Algoritma Linear Regression", *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 8, No. 1, pp. 41-50, 2021. ISSN: 2407-4322.
- [2] Victor Marudut Mulia Siregar, "Perancangan Aplikasi Data Mining Untuk Memprediksi Penjualan Menggunakan Metode Decision Tree Pada Apotik Ths Pematangsiantar", *J. Murni Politeknik Bisnis Indonesia*, vol. 7, no. 1, pp. 51-61, 2017. ISSN: 2338-8196.
- [3] Arif Fadilah, "Prediksi Harga Rumah di Kota Bandung Bagian Timur dengan Menggunakan Metode Moving Average", *Jurnal Teknik Informatika*, vol.7 no. 2, pp. 40-49, 2020. ISSN: 2355-9365.
- [4] Luluk Suryani dan Kasmi Edy, "Application Development "Lost & Found" Android Based Using Term Frequency - Inverse Document Frequency (Tf-Idf) And *Cosine Similarity* Method", *J. Electro Luceat*, vol. 6, no. 2, Mei, pp. 1-15, 2020.
- [5] Green Arther Sandag, "Prediksi Rating Aplikasi App Store Menggunakan Algoritma Random Forest", *Jurnal Ilmu Komputer*, vol. 6, no. 2, pp. 167-177, 2020. ISSN: 2541-2221.
- [6] Maryati Puji Lestari, Deden Jacob Witarasyah, Faqih Hamami, "Peramalan Pertambahan Pasien COVID-19 Menggunakan Support Vector Regression", *Jurnal Teknologi Informatika*, vol. 8, no. 5, pp. 94-106, 2021. ISSN: 2355-9365.
- [7] Feri Irawan, Sumijan dan Yuhandri, "Prediksi Tingkat Produksi Buah Kelapa Sawit dengan Metode Single Moving Average", *Jurnal Informasi dan Teknologi*, vol. 3, no.4, Juli, pp. 251-256, 2021. ISSN: 2714-9730.
- [8] A. J. Syahid dan D. Mahdiana, "Perbandingan Algoritma Untuk Klasifikasi Analisis Sentimen Terhadap GeNose Pada Media Sosial Twitter", *SemanTIK*, vol. 7, no. 1, pp. 9-16, 2021. <https://doi.org/10.5281/zenodo.5034916>.
- [9] Heri Setyawan, Sri Hariyati Fitriasih dan Retno Tri Vlandari, "Prediksi Tingkat Produksi Buah Kelapa Sawit dengan Metode Single Moving Average", *J.TIKomSiN*, vol. 9, no. 2, pp. 1-10, 2021. ISSN: 2338-4018 <https://doi.org/10.30646/tikomsin.v9i2.53>.
- [10] Narkhede, Sarang. 2018. "Understanding Confusion Matrix." *Towards Data Science*, via Medium.
- [11] Pardomuan Robinson Sihumbing dan Ade Marsinta Arsani, "PERBANDINGAN METODE MACHINE LEARNING DALAM KLASIFIKASI KEMISKINAN DI INDONESIA TAHUN 2018", *Jurnal Teknik Informatika*, Vol. 2, No. 1, pp. 51-56, 2021. ISSN: 2723-3871.