Terbit online pada laman web jurnal: **http://journal.isas.or.id/index.php/JACOST**

JOURNAL OF APPLIED COMPUTER SCIENCE AND TECHNOLOGY (JACOST)

Vol. 4 No. 1 (2023) 58 – 64 | ISSN: 2723-1453 (Media Online)

Analisis Perbandingan Metode Regresi Linier, Random Forest Regression dan Gradient Boosted Trees Regression Method untuk Prediksi Harga Rumah

Evita Fitri

[1]Program Studi Sistem Informasi, Fakultas Teknologi Informasi, Universitas Nusa Mandiri

[1]evita.etv@nusamandiri.ac.id

Abstract

The need for a place to live is one that many people prepare, both millennials and adults and the elderly. With the continued increase in population growth in Indonesia and increasing public interest in buying a place to live early on, this can make not all groups of people have a place to live or a house that is quite livable. Related to this, the public needs up-to-date information related to predictions of house prices both for housing and second-hand housing prices for planning purposes in the future. The purpose of this study is to carry out a comparative analysis of the prediction results of house prices with several Machine Learning algorithms consist of Linear Regression, Random Forest Regression and Gradient Boosted Trees Regression. Evaluation for all the method applying Cross-Validation. The evaluation is seen from the smallest Root Mean Square Error (RMSE) error rate of each testing method. The results of this study are the Random Forest Regression obtained an RMSE value of 0.440, the Linear Regression model obtained an RMSE value of 0.515 and the RMSE value of Gradient Boosted Trees Regression of 0.508. The results were obtained from testing a dataset of 2011 records with a division of 80% for data training and 20% for data testing, the data has 6 attributes used in testing including house prices, land area, building area, number of bathrooms, number of bedrooms and the number of garages. In this study, prediction results using the Random Forest Regression method yielded the highest accuracy of 81.5% compared to the Linear Regression and Gradient Boosted Trees Regression methods.

Keywords: Prediction, Machine Learning, Linear Regression, Random Forest Regression, Gradient Boosted Trees Regression, Analysis, Comparison.

Abstrak

Kebutuhan akan tempat tinggal menjadi salah satu yang banyak orang-orang persiapkan, baik dari kalangan milenial ataupun dewasa dan orang tua. Dengan terus bertambahnya pertumbuhan penduduk di Indonesia dan bertambahnya minat masyarakat dalam membeli tempat tinggal sejak dini, hal tersebut dapat membuat tidak semua golongan dari masyarakat untuk memiliki tempat tinggal atau rumah yang cukup layak huni. Terkait hal tersebut, masyarakat membutuhkan informasi *uptodate* terkait prediksi harga rumah untuk keperluan perencanaan dimasa yang akan datang. Adapun tujuan dari penelitian ini yaitu melakukan analisa perbandingan hasil prediksi harga rumah dengan beberapa algoritma *Machine Learning* yaitu *Linear Regression, Random Forest Regression* serta *Gradient Boosted Trees Regression.* Dan menerapkan *Cross-Validation* pada setiap pengujian beberapa metode tersebut. Evaluasi dilihat dari *error rate Root Mean Square Error* (RMSE) terkecil dari setiap pengujian metode. Hasil dari penelitian ini yaitu *pada Random Forest Regression* didapat nilai RMSE 0,440, pada model *Linear Regression* didapat nilai RMSE 0,515 dan pada *Gradient Boosted Trees Regression* nilai RMSE sebesar 0,508. Adapun hasil tersebut didapat dari pengujian dataset sebesar 2011 *record* dengan pembagian 80% untuk data *traning* dan 20% untuk data *testing*, data tersebut memiliki 6 atribut yang digunakan dalam pengujian diantaranya harga rumah, luas tanah, luas bangunan, jumlah kamar mandi, jumlah kamar tidur dan jumlah garasi. Pada penelitian ini, hasil prediksi menggunakan metode *Random Forest Regression* menghasilkan akurasi tertinggi sebesar 81,5% dibandingkan dengan metode *Linear Regression* dan *Gradient Boosted Trees Regression*.

Kata kunci: Prediksi, *Machine Learning, Linear Regression, Random Forest Regression, Gradient Boosted Trees Regression,* Analisa, Perbandingan.

1. Pendahuluan

Pergerakan pertumbuhan ekonomi pada setiap negara pada umumnya memungkinkan berbeda-beda, termasuk pada prediksi terkait pertumbuhan ekonomi pasca dilandanya banyak negara oleh pandemi Covid 19, adapun pada saat ini Indonesia sendiri merupakan salah satu negara yang telah memulai bangkit kembali dalam memajukan perekonomian pasca dilandanya pandemi Covid 19, dengan hasil ramalan terkait prediksi

pertumbuhan ekonomi Indonesia terbilang masih mengalami resesi atau mengalami pertumbuhan negatif dengan nilai prediksi sebesar 0,000242 pada nilai error *Mean Squared Error* (MSE) dan nilai *Mean Average Error* (MAE) nya. [1]. Berkaitan dengan pertumbuhan ekonomi suatu negara, hal ini juga berpengaruh pada status finansial bagi setiap individu di negara tersebut, dimana pada saat ini, finansial merupakan salah satu aspek yang mempengaruhi keberlanjutan kehidupan seseorang diantaranya adalah memiliki tempat tinggal yang nyaman.

Rumah atau tempat tinggal merupakan kebutuhan dasar bagi setiap orang yang kumulatifnya setiap tahun semakin banyak dibutuhkan karena perkembangan penduduk Indonesia yang semakin bertambah, hal ini menjadikan bahwa setiap orang sadar akan pentingnya dalam memiliki sebuah tempat tinggal berupa rumah [2].

Adapun hal ini ditunjukkan bahwa bukan hanya orang yang lebih dewasa saja namun para generasi milenial atau setiap penduduk Indonesia yang mulai beranjak dewasa pun ikut berlomba dalam memilih dan mulai mempunyai rumah untuk keperluan pribadinya [3]. Terkait hal tersebut, terdapat faktor-faktor yang mempengaruhi dan menjadi bahan pertimbangan para generasi milenial dalam memilih rumah diantaranya ialah faktor internal yang meliputi pendidikan, pekerjaan hingga pendapatan orang tersebut, lalu faktor keluarga, dan faktor eksternal yang meliputi fasilitas kredit serta kebijakan-kebijakan yang ada terkait tempat tinggal.

Penelitian [3] menyatakan bahwa faktor yang menjadi bahan pertimbangan generasi milenial dalam mulai memiliki rumah ialah faktor keuangan dan lingkungan , keluarga dan eksternal, kondisi fisik dari rumah tinggal, internal dan selera dari masing-masing individu. Sedangkan kendala finansial utama yang terjadi pada generasi milenial, tidak tersedianya dukungan finansial dari orang terdekat sehingga ragu melakukan pembelian rumah.

Penelitian berikutnya menjelaskan bahwa index harga rumah merupakan variabel untuk memperkirakan ketidak konsistenan sebuah harga rumah, hal ini karena harga rumah yang berkorelasi erat dengan beberapa faktor seperti lokasi, kota serta populasinya sehingga dilakukannya penelitian dalam prediksi harga rumah. Penelitian ini menggunakan teknik *machine learning Random Forest* dengan dataset berisi 500 data dan 14 fitur, adapun hasil dari penelitian ini perbandingan harga prediksi yang didapat dan harga aktual mengungkapkan bahwa model tersebut memiliki nilai prediksi yang dapat diterima jika dibandingkan dengan nilai aktual dengan margin kesalahan ±5 [4].

Tujuan dari penelitian ini ialah melakukan analisa perbandingan dari beberapa algoritma model *time series* yang digunakan untuk memprediksi nilai akurasi harga suatu rumah.

Penelitian selanjutnya melakukan prediksi harga rumah dengan mengambil dataset dari data rumah di Bandar Lampung yang berisi fitur harga, lokasi serta spesifikasi bangunan rumah, adapun hasil dari menentukan akurasi pada prediksi rumah yang telah dilakukan yaitu model *Polinomial Regresi* kernel mencapai R^2 tertinggi yaitu kernel linier 95,99% dan kernel gaussian mencapai R^2 masing-masing 90,99% dan 81,43%. Sedangkan pada uji coba model klasifikasi akurasi diperoleh pada 8 kelas kernel gaussian sebesar 91,18%, dan kernel linier dan kernel polinomial mendapatkan akurasi sebesar 90,20% dan 89,90% [5].

Selanjutnya penelitian [6] menerapkan metode *Gradient Boosted Trees* pada pengujian prediksi harga rumah yang menghasilkan akurasi 90,00% serta penelitian [7] melakukan prediksi dengan membagi dataset menjadi 80% - 20 % dan menerapkan komparasi model dengan hasil akhir metode *Random Forest Regression* yang mendapatkan hasil akurasi yang cukup baik sebesar 81,6% .

Analisa prediksi terkait harga rumah juga dilakukan pada penelitian [8] hal ini dilakukan berdasarkan kesesuaian spesifikasi dari index harga rumah, dengan metode *Multiple Linear Regression* dihasilkan bahwa 66% nilai akurasi didapatkan berdasarkan pengujian 1001 data set dengan jumlah fitur sebanyak 7.

Ditinjau dari beberapa kajian tersebut, pada penelitian ini dilakukan komparasi algoritma untuk memprediksi index harga rumah berdasarkan faktor-faktor yang mempengaruhi harga rumah di Indonesia dengan metode *Random Forest Regression,, Linear Regression* dan *Gradient Boosted Trees Regression*, serta dilakukan ujicoba *K-Fold Cross Validation* pada setiap pengujian metodenya. Adapun pada penelitian ini, digunakan data yang bersumber dari kaggle (https://www.kaggle.com/wisnuanggara/daftar-harga-rumah) yang merupakan kumpulan data index harga rumah wilayah Jakarta, Indonesia dengan jumlah data sebanyak 2011 *record*.

Random Forest Regression merupakan algoritma *machine learning* yang proses pengujiannya menggunakan konsep *supervised* dalam membangun kelas *classifier*. Algoritma ini mengkombinasikan prediksi berdasarkan *Multiple Decision Tree* [9].

Sedangkan Metode *Gradient Boosted Trees Regression* merupakan metode penggabungan dari beberapa *decision trees*, tetapi tidak seperti *Random Forest*, di mana *tree* dilatih secara paralel pada sampel *bootstrap* dari kumpulan data asli, *tree* dilatih secara berurutan, metode ini juga merupakan metode gradient descent yang dimodifikasi dengan *boosting algorithm* untuk meningkatkan akurasi [10].

Linear Regression sendiri merupakan jenis model dengan teknik yang bertujuan dalam menganalisis estimasi nilai variabel dependen dengan rentang nilai

DOI : https://doi.org/10.52158/jacost.491

variabel independen [11]. Secara umum algoritma ini terbagi menjadi 2 yaitu *Simple Linear Regression* dan *Multiple Linear Regression. Simple Linear Regression* merupakan hubungan antar satu variabel dependen dan satu variabel independen, sedangkan *Multiple Linear Regression* merupakan hubungan antar satu variabel dependen dan dua atau lebih variabel independennya [12].

Pada penelitian ini, dalam menerapkan metode *Linear Regression*, peneliti menggunakan *Multiple Linear Regression* dengan rincian dataset yang terdiri dari satu variabel dependen (harga rumah) dan lima variabel independen yaitu luas bangunan, luas tanah, jumlah kamar mandi, kamar tidur dan garasi.

2. Metode Penelitian

2.1. Sumber Data

Pada penelitian ini, sumber data yang digunakan yaitu menggunakan data sekunder yang bersifat *time series*, adapun data ini merupakan data histori harga-harga rumah yang didapat atau bersumber dari kaggle bersifat *open source*.

Dataset yang digunakan terdapat 2 bagian yang berbeda wilayah dengan jumlah 1010 record pada wilayah A dan 1001 record pada wilayah B dan masing-masing terdapat 7 fitur. Harga rumah didapat dari proses pengumpulan data harga di beberapa *website* penjualan rumah, adapun sampel harga yang diambil ialah harga rumah di daerah tebet dan Jakarta Selatan. Berikut detail dataset yang digunakan.

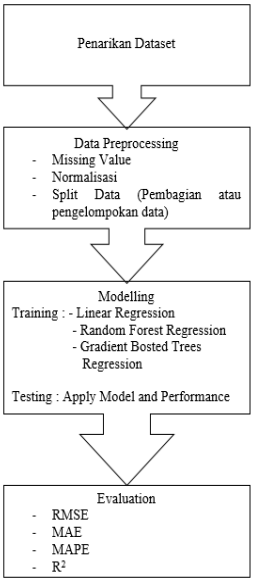
Tabel 1. Detail Dataset Penelitian

No	Harga Rumah (Wilayah)	Source Dataset	Atribut	Jumlah Instance
1	Harga Rumah Jakarta Selatan	https://www.kaggle.com/wisnuanggara/daftar-harga-rumah	Harga, Luas Tanah, Luas Bangunan, Kamar Tidur, Kamar Mandi, Garasi, Kota	1001
2	Harga Rumah Khusus Hanya Wilayah Tebet	https://www.kaggle.com/wisnuanggara/daftar-harga-rumah	Nama Rumah, Harga, Luas Bangunan, Luas Tanah, Kamar Tidur, Kamar Mandi, Garasi	1010

2.2. Metode Penelitian

Pada penelitian ini, peneliti menggunakan salah satu metodologi standar dalam data mining yaitu metode CRISP-DM yaitu model *Cross Standard Industry For Data Mining* dengan tahapan yang dilakukan dimulai pada analisa *business understanding*, lalu data understanding yang dilanjutkan dengan preparation data serta *modelling* dan diakhiri sampai dengan tahapan evaluasi.

Gambar 1 yang menunjukkan proses tahapan yang dilakukan pada penelitian ini.



Gambar 1. Tahapan Metode Penelitian

Pada Gambar 1 memperlihatkan alur tahapan pada metode penelitian yang dilakukan diantaranya :

1. Penarikan dataset, Pada penelitian ini, dilakukan penarikan data sebagai pengujian beberapa metode machine learning, adapun data yang digunakan adalah dua jenis data harga rumah berbeda lokasi dengan total dataset sebanyak 1010 dan 1001 dataset dan 7 fitur.
2. Data *Preprocessing*, Pada tahap ini dilakukan proses pengolahan data diantaranya pengecekan *missing value* data yang pada pengerjaannya yaitu membersihkan data dari data yang kosong atau hilang, selanjutnya dilakukan normalisasi data sebagai proses data transformation untuk menyeimbangkan nilai pada setiap *record*, dikarenakan setiap data memiliki rentang data yang berbeda, dan terakhir yaitu adanya proses pembagian data yang terdiri dari data training dan data testing.
3. Pengujian Model Prediksi, Selanjutnya yaitu tahap *modelling* untuk memprediksi data hasil *preprocessing*, pada tahap ini prediksi dilakukan dengan membandingkan tiga model prediksi yaitu *Linear Regression, Random Forest Regression* dan *Gradient Boosted Trees Regression.*
4. Evaluation, Dan tahapan terakhir yaitu proses evaluasi, adapun tahapan evaluasi yang dilakukan pada penelitian ini yaitu menggunakan tools KNIME, dimana validasi data dilakukan pada data *training* dengan nilai 10 *K-Fold Cross-Validation.* Pada proses evaluasi ini diukur pada hasil yang diperoleh dari nilai R^2, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) dan Mean Absolute Percentage Error (MAPE) yang lebih

DOI : https://doi.org/10.52158/jacost.491

rendah dari masing-masing uji coba dataset dengan tiga model prediksi, selanjutnya dibandingkan dan didapat hasil kesimpulan model prediksi mana yang cukup baik digunakan.

2.3. Metode Pengujian Dataset

Penelitian ini melakukan perbandingan dengan menggunakan tiga metode algoritma, diantaranya metode Multiple Linear Regression dengan bentuk umum persamaan sebagai berikut [13] :

$$Y = \beta 0 + \beta\ 1X1\ \beta\ 2X2 + \ldots\ldots\ldots + \beta\ nXn \quad (1)$$

Dengan Y=prediksi nilai variabel dependen, X= variabel *independent,m,* β0= konstanta, βn= Bobot (koefisien) regresi untuk variabel *independent*

Selanjutnya penelitian [7] juga menyimpulkan bahwa *Random Forest Regression* mendapatkan hasil yang lebih akurat dibandingkan dengan linear regression, adapun persamaan pada metode ini sebagai berikut [14]:

$$\hat{Y}i = \frac{1}{N_{tree}} \sum_{n=1}^{N_{tree}} \hat{Y}n \quad (2)$$

Dengan $\hat{Y}i$ = Hasil prediksi, N_{tree} =Total nilai pohon, $\hat{Y}n$= Hasil prediksi pohon ke-n

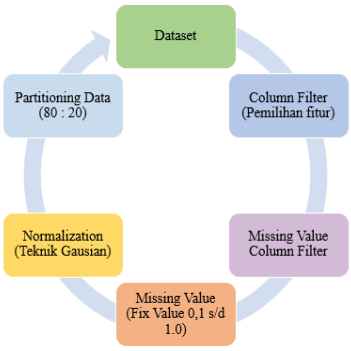
Dan perbandingan lainnya juga dilakukan dengan metode *Gradient Boosted Trees Regression* dengan persamaan berikut:

$$f_M(x_i) = \sum_{m=1}^{M} \eta . h_m(x_i; q_m) \quad (3)$$

Dimana $h_m(x_i; q_m$ adalah merupakan nilai pohon keputusan dan η merupakan parameter yang menentukan seberapa cepat ditemukannya hasil dari metode [6].

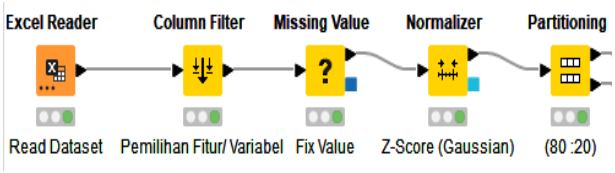
2.4. Metode Pengolahan Dataset

Adapun pada pengolahan dataset penelitian, seperti yang dideskripsikan pada Gambar 2, pada pengolahan dataset dilakukan pengecekan awal serta *preprocessing* data, hal ini diperuntukan guna membersihkan data yang tidak sempurna atau *missing value* serta data yang tidak valid, selain dilakukannya pengecekan dataset sebelum nantinya digunakan dalam ujicoba pemodelan, dilakukan juga tahapan normalisasi pada seluruh dataset.



Gambar 2. Tahapan Pengolahan Dataset Penelitian

Pada Gambar 2 menunjukan beberapa tahapan dalam mengolah dataset penelitian yang dilakukan oleh peneliti, tahapan pengolahan dataset yang dilakukan diantaranya:



Gambar 3. Visualisasi Preprocessing Data

1. Pada proses awal digunakan fungsi *file reader*, proses ini dilakukan sebagai tahap awal pembacaan file dataset
2. Proses *column filter* digunakan dalam proses *preprocessing* yaitu sebagai proses penyaringan atau pemilihan fitur yang akan digunakan dalam uji coba. Pada penelitian ini, terdapat attribute yang tidak digunakan yaitu attribute kota, sehingga attribute yang digunakan hingga akhir yaitu terdapat 5 fitur.
3. Missing value *column filter* ini digunakan sebagai proses pembersihan khusus untuk seluruh data yang atributnya telah dipilih atau digunakan.
4. Sedangkan pada *Missing Value* lanjutan digunakan sebagai proses pembersihan data dari data *record* yang tidak lengkap atau atribut yang tidak relevan untuk seluruh dataset.
5. Normalisasi, proses normalisasi ini dimaksudkan untuk menyeimbangkan nilai antara fitur-fitur yang ada, teknik yang digunakan pada penelitian ini ialah teknik *Gaussian*.
6. *Partitioning* atau *split* data pada pengolahan dataset dilakukan dengan jumlah pembagian data *training* dan data *testing* masing-masing sebesar 80 : 20 dengan teknik *draw randomly*.

3. Hasil dan Pembahasan

3.1. Dataset Penelitian

Data harga rumah yang digunakan didapatkan dari *website* kaggle, digunakan 2 sub data harga rumah, yaitu pada wilayah Jakarta Selatan dan data harga rumah khusu daerah Tebet. Adapun jumlah keseluruhan data yaitu 2011 data.

Tabel 2. *Sample* Dataset Harga Rumah Wilayah Jakarta Selatan

Harga	Lt	Lb	Jkt	Jkm	Grs	Kota
4.900.000.000	251	300	5	4	Ada	Jaksel
28.000.000.000	1340	575	4	5	Ada	Jaksel
10.000.000.000	460	300	4	4	Ada	Jaksel
670.000.000	70	69	3	2	Tidak Ada	Jaksel
480.000.000	66	42	2	1	Tidak Ada	Jaksel

Tabel 2 merupakan detail terkait data harga rumah di wilayah Jakarta Selatan dengan fitur diantaranya harga rumah, luas tanah, luas bangunan, jumlah kamar tidur, jumlah kamar mandi, jumlah garasi dan nama kota,

DOI : https://doi.org/10.52158/jacost.491

adapun jumlah *record* pada harga rumah wilayah Jakarta Selatan yaitu 1001 record.

Tabel 3 *Sample* Dataset Harga Rumah Pada Wilayah Tebet

Nama Rumah	Harga	Lb	Lt	Kt	Km	Grs
Rumah Mewah 2 Lantai Hanya 3 Menit Ke Tebet, Tebet, Jakarta Selatan	3000000000	267	250	4	4	4
Rumah lama di Tebet, dekat MT Haryono dan tol dalam kota, jalan 1 mobil hanya 30M dr jln besar, Tebet, Jakarta Selatan	2600000000	120	150	3	2	1
Rumah Bagus Keren Jalan Lebar Di Area & Kawasan Terbaik Tebet, Tebet, Jakarta Selatan	1050000000	350	247	4	4	0
Minimalis Baru Jalan 1 Mobil Akses Mudah Dekat ke Jalan Lebar, Tebet, Jakarta Selatan	3250000000	125	90	3	3	0
Minimalis Baru Jalan 2 Mobil Tebet Timur, Tebet, Jakarta Selatan	4500000000	250	96	5	4	1

Tabel 3 merupakan detail dataset penelitian yang digunakan dalam pengujian model data mining, pada sampel diatas terdapat 7 fitur yaitu nama rumah, harga, luas bangunan, luas tanah, jumlah kamar tidur, jumlah kamar mandi dan jumlah garasi dengan total instance 1010 record.

Tabel 4. Hasil Normalisasi Fitur

Row id	Harga	LB	LT	KT	KM	GRS
0	-0,522	-0,318	-0,097	-1.061	-0,428	-1,271
1	-0,413	-0,543	-0,558	-0,425	-0,428	0.052
2	-0,631	-0,054	0.07	-0,425	0,276	1,376
3	-0,981	-1,33	-1,18	-1,697	-1,132	-1,271
4	0,187	0,694	0.653	0,847	0,98	0,714
5	-0,362	0.132	-0,464	0,211	-0428	0,714
6	-0,685	-0,88	-0,486	-1,061	-1,132	-0,609

Tabel 4 menunjukkan hasil dataset harga rumah yang telah dilakukan proses normalisasi pada setiap fiturnya, adapun data yang telah dinormalisasi di split sebagai data training 80 % dengan jumlah 808 data dan data testing sebesar 202 data.

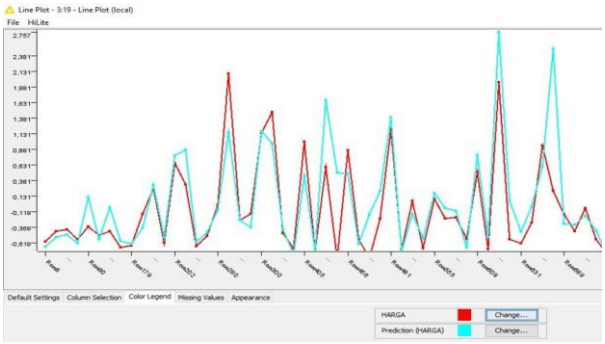
3.2. Pengujian Model Prediksi

Pada proses pengujian model diantaranya dilakukan split data dan pemodelan *Linear Regression, Random Forest Regression* dan *Gradient Boosted Trees Regression.* dengan menerapkan *K-Fold Cross-Validation.* Adapun batasan dari penelitian ini yaitu mengolah dataset dengan mengujikan beberapa metode untuk menghasilkan model prediksi dengan akurasi tertinggi dan mengevaluasi hasil *error rate* pada setiap pengujian metodenya

3.2.1. Pengujian model prediksi *Linear Regression* dengan *Partitioning* dan Cross-Validation

Pada pengujian ini, dua dataset masing-masing diuji dengan model dan proses yang sama, hasil dari processing dataset dilakukan pembagian data atau partitioning dengan jumlah 80 : 20 untuk masing-masing data training dan data testing dan setelahnya dilakukan validasi *Cross-Validation* pada setiap model pengujian,

Pada data training dilakukan pengujian model *Linear Regression Learner* sebagai data yang dijadikan sampel pemodelan dengan menggunakan K-10 *Cross-Validation* pada *class column* (atribut harga) dengan teknik *Stratified sampling*. Begitupun dengan data *testing* dilakukan pula pengujian model dengan membaca hasil dari pengujian data *training* yang sebelumnya dilakukan, adapun berikut visualisasi terhadap harga awal dataset dan hasil dari prediksi harga rumah yang telah diujikan dengan model prediksi *Linear Regression*:



Gambar 4. *Line Plot* Prediksi Harga Rumah *Linear Regression*

Pada Gambar 4 menunjukkan hasil prediksi *class column* yaitu harga pada data tersebut dengan *line red* atau garis merah merupakan harga awal rumah yang digunakan pada awal pengujian, harga ini merupakan harga awal dari dataset yang digunakan, dan pada *line blue* merupakan harga prediksi yang dihasilkan oleh pemodelan yang telah peneliti jelaskan sebelumnya, yaitu menggunakan *Linear Regression* dengan *Cross-Validation K-10*. Adapun didapat diantaranya total nilai *score error* pada model ini terlihat pada Tabel 5.

Tabel 5 menunjukkan perbandingan prediksi untuk pengujian dataset harga rumah pada daerah tebet dengan pembagian data 80: 20, jumlah dara *training* 808 *intacnce* dan data *testing* 202 *instance* dengan hasil evaluasi *error* yang cukup kecil 0.569 pada *training* dan 0.433 pada *testing*, adapun pada pengujian dataset harga rumah di daerah jakarta dengan jumlah data *training* 800 *instance* dan *testing* sebesar 201 *instance*, hasil dari evaluasi *error* yang didapat ialah diangka 0.663 dan 0.768. Dari kedua pengujian tersebut pemodelan dilakukan dengan cara yang sama namun dengan banyaknya jumlah data atau record yang digunakan

DOI : https://doi.org/10.52158/jacost.491

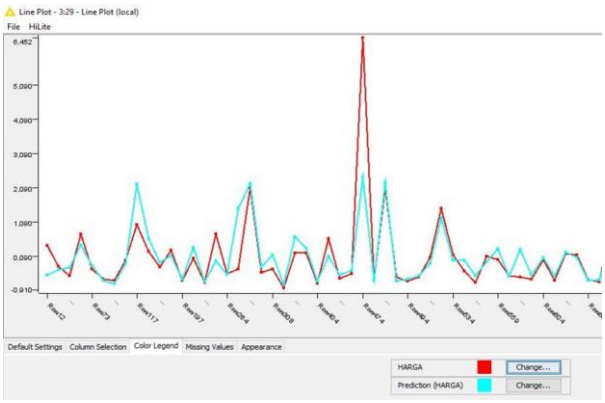
sebagai pemodelan, hal ini dapat mempengaruhi hasil prediksi *error* yang didapat dari model *Linear Regression.*

Tabel 5. Hasil Prediksi Menggunakan *Linear Regression*

Result Evaluation	Dataset Harga Rumah Tebet		Dataset Harga Rumah Jakarta Selatan	
	Training (808)	Testing (202)	Training (800)	Testing (201)
R^2	0.690	0.747	0.548	0.460
Mean absolute error	0.300	0.271	0.292	0.345
Mean squared error	0.306	0.265	0.440	0.590
Root mean squared error	0.553	0.515	0.663	0.768
Mean absolute percentage error	1.117	1.014	3.839	1.611
Adjusted R^2	0.690	0.747	0.548	0.460

3.2.2. Pengujian model prediksi *Random Forest Regression* dengan *Partitioning* dan *Cross-Validation*

Pengujian model *Random Forest Regression* dilakukan pada kedua dataset, fitur harga sebagai target kolom pada pengujian *Random Forest Learner*, adapun pada pengujian model ini jumlah *number of model forest* ditetapkan 100 serta jumlah k *Cross-Validation* sebesar 10. Pada pembuatan model digunakan data *training* sebagai pemodelan dan *testing* yang diujikan menggunakan hasil pemodelan *training* tersebut. Adapun hasil visualisasi prediksi dari salah satu pengujian dataset pada model ini yaitu:



Gambar 5..*Line Plot* Prediksi Harga Rumah (Case Dataset Harga Rumah Tebet) Model *Random Forest Regression*

Pada Gambar 5 menunjukkan *line blue* sebagai nilai prediksi dan *line red* sebagai nilai harga awal, dimana pada *line plot* tersebut menunjukkan beberapa kesinkronan antara prediksi dengan nilai awal, namun tetap terdapat beberapa bagian yang *miss* dengan hasil ketidaksesuaian antara nilai fitur awal dengan hasil prediksi. Jika dilihat dari kedua desain *line plot* (Gambar 4 dan Gambar 3) dapat terlihat bahwa prediksi pada design *line plot* Gambar 6 model *Random Forest Regression* lebih menunjukkan kesesuaian atau ketepatan terhadap nilai awal dan hasil prediksi dibandingkan dengan design *line plot* hasil prediksi menggunakan model *Linear Regression* .

Tabel 6. Hasil Prediksi Menggunakan *Random Forest Regression*

Result Evaluation	Dataset Harga Rumah Tebet		Dataset Harga Rumah Jakarta Selatan	
	Training (808)	Testing (202)	Training (800)	Testing (201)
R^2	0.740	0.815	0.631	0.473
Mean absolute error	0.243	0.220	0.262	0.302
Mean squared error	0.257	0.194	0.360	0.575
Root mean squared error	0.507	0.440	0.600	0.758
Mean absolute percentage error	0.898	0.732	2.876	1.574
Adjusted R^2	0.740	0.815	0.631	0.473

Tabel 6 merupakan hasil pengujian *error rate* menggunakan model *Random Forest Regression*, pada hasil tersebut nilai RMSE terkecil ada pada pengujian *testing* yaitu sebesar 0.433, jika dilihat dengan Tabel 4 hasil prediksi *error* menggunakan *Linear Regression* didapat hasil yang cukup tepat terhadap nilai *error* RMSE pada pengujian *testing* dataset harga rumah tebet. Namun rata-rata hasil dari pengujian model prediksi menggunakan *Random Forest Regression* ini cukup signifikan kecil pada nilai errornya.

3.2.3. Pengujian model prediksi *Gradient Boosted Trees Regression* dengan *Partitioning* dan *Cross-Validation*

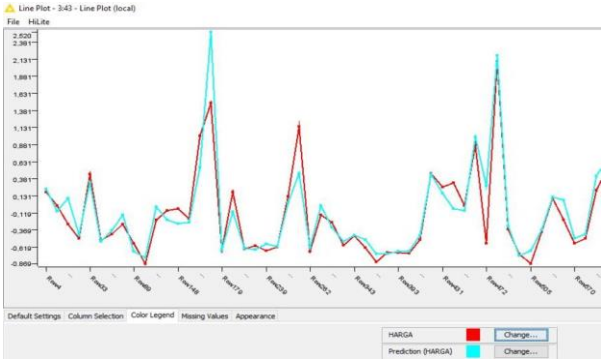
Pada pengujian metode *Gradient Boosted Trees Regression*, menggunakan limit number dari *tree* nya sebesar 4 dengan jumlah parameter *learning rate* 0,1. Pengujian learning rate dengan jumlah tersebut merupakan nilai *rate* yang lebih optimal yang dicoba oleh peneliti untuk menghitung nilai koreksi bobot pada waktu proses training data.

Tabel 7. Hasil Prediksi Menggunakan *Gradient Boosted Trees Regression*

Result Evaluation	Dataset Harga Rumah Tebet		Dataset Harga Rumah Jakarta Selatan	
	Training (808)	Testing (202)	Training (800)	Testing (201)
R^2	0.702	0.754	0.329	0,209
Mean absolute error	0.247	0.224	0.249	0,350
Mean squared error	0.294	0.258	0.351	0,864
Root mean squared error	0.543	0.508	0.592	0.930
Mean absolute percentage error	0.804	0.715	2.681	1.232
Adjusted R^2	0.702	0.754	0.329	0.209

DOI : https://doi.org/10.52158/jacost.491

Pada hasil tabel 7, *Gradient Boosted Trees Regression* menghasilkan cukup tinggi akurasi 75% pada training dengan nilai *error* yang cukup kecil dibandingkan dengan metode *Linear Regression* yakni sebesar 0.508 pada nilai RMSE nya.



Gambar 7 *Line Plot* Prediksi Harga Rumah (Case Dataset Harga Rumah Tebet) Model *Gradient Boosted Trees Regression*

Pada hasil visualisasi gambar 7, metode *Gradient Boosted Trees Regression* memang menghasilkan cukup baik hasil akurasi dibandingkan dengan metode lain seperti *Linear Regression*, hal ini selaras dengan penelitian sebelumnya yang didapat hasil akurasi sebesar 90%. Namun pada pengujian penelitian yang dilakukan oleh peneliti didapatkan hasil yang berbeda, yaitu pengujian metode *Random Forest Regression* dapat menghasilkan nilai *error* yang lebih kecil dan nilai akurasi yang cukup besar dibandingkan dengan *Gradient Boosted Trees Regression* serta *Linear Regression.* Adapun pada bagian pembahasan serta hasil yang peneliti lakukan dibatasi hingga pengujian yang menghasilkan akurasi dan analisa nilai *error* pada setiap metedo yang digunakan.

4. Kesimpulan

Berdasarkan pengujian metode *Linear Regression*, *Random Forest Regression* dan *Gradient Boosted Trees Regression* pada dataset harga rumah, telah didapat hasil perbandingan prediksi dengan *error rate* lebih kecil pada metode *Random Forest Regression*. Evaluasi prediksi dilakukan dengan melihat hasil *error* pada RMSE setiap model, yaitu pada *Random Forest Regression* didapat nilai *error* 0.440, *Linear Regression* didapat nilai *error* terkecil di range 0.515 dan pada *Gradient Boosted Trees Regression* sebesar 0.508. Adapun pada pengujian setiap metode, jumlah data cukup berpengaruh terhadap hasil prediksi, semakin banyak data yang diujikan sebagai pemodelan *learner* setiap metode, maka hasil tersebut dapat dikatakan lebih akurat. Jumlah data yang banyak juga harus dilakukan *preprocessing* sebagai pendukung dari pengujian metode-metode lainnya. Adapun hasil-hasil tersebut dapat dievaluasi kembali dengan menambahkan beberapa teknik serta penambahan jumlah dataset sebagai pengujian model algoritma lainnya, sehingga nantinya akan menghasilkan pengujian prediksi yang lebih akurat. Pada penelitian ini nilai akurasi prediksi harga rumah menggunakan metode *Random Forest Regression* menghasilkan akurasi tertinggi sebesar 81,5% dibandingkan dengan metode *Linear Regression* dan *Gradient Boosted Trees Regression.*

Daftar Rujukan

[1] A. Asrirawan, S. U. Permata, and M. I. Fauzan, "Pendekatan Univariate Time Series Modelling untuk Prediksi Kuartalan Pertumbuhan Ekonomi Indonesia Pasca Vaksinasi COVID-19," Jambura J. Math., vol. 4, no. 1, pp. 86–103, 2022.

[2] R. P. Sari and L. Novitasari, "Sistem Penentuan Kelayakan Kredit Pemilikan Rumah Non-Subsidi Menggunakan Metode Weight Product," J. Rekayasa Teknol. Inf., vol. 6, no. 1, p. 18, 2022.

[3] D. D. Wijaya and N. Anastasia, "Pertimbangan Generasi Milenial Pada Kepemilikan Rumah dan Kendala Finansial," J. Manaj. Aset dan Penilai., vol. 1, no. 2, pp. 11–20, 2021.

[4] A. B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, and G. Oluwadara, "House Price Prediction using Random Forest Machine Learning Technique," Procedia Comput. Sci., vol. 199, pp. 806–813, 2021.

[5] F. R. Lumbanraja, R. A. Saputra, K. Muludi, A. Hijriani, and A. Junaidi, "Implementasi Support Vector Machine Dalam Memprediksi Harga Rumah Pada Perumahan Di Kota Bandar Lampung," J. Pepadun, vol. 2, no. 3, pp. 327–335, 2021.

[6] A. Hjort, J. Pensar, I. Scheel, and D. E. Sommervoll, "House price prediction with gradient boosted trees under different loss functions," J. Prop. Res., vol. 39, no. 4, pp. 338–364, 2022.

[7] C. Haryanto, N. Rahaningsih, F. M. Basysyar, K. Cirebon, R. F. Regression, and H. Rumah, "Komparasi Algoritma Machine Learning Dalam Memprediksi Harga Rumah," vol. 1, no. 1, pp. 533–539, 2023.

[8] M. L. Mu'tashim, T. Muhayat, S. A. Damayanti, H. N. Zaki, and R. Wirawan, "Analisis Prediksi Harga Rumah Sesuai Spesifikasi Menggunakan Multiple Linear Regression," Inform. J. Ilmu Komput., vol. 17, no. 3, p. 238, 2021.

[9] S. Saadah and H. Salsabila, "Prediksi Harga Bitcoin Menggunakan Metode Random Forest (Studi Kasus: Data Acak Pada Awal Masa Pandemic Covid-19)," J. Komput. Terap., vol. 7, no. Vol. 7 No. 1 (2021), pp. 24–32, 2021.

[10] J. Athalia et al., "Perbandingan Analisis Faktor Penentu Penjualan PT . X Menggunakan LASSO Regression dan Gradient Boosted Regression Tree," J. Infra, vol. 9, no. 1, 2021.

[11] Mikhael, F. Andreas, and U. Enri, "Perbandingan Algoritma Linear Regression, Neural Network, Deep Learning, Dan K-Nearest Neighbor (K-Nn) Untuk Prediksi Harga Bitcoin," J. Sist. Inf., vol. 14, no. 1, pp. 2450–2464, 2022.

[12] H. W. Herwanto, T. Widiyaningtyas, and P. Indriana, "Penerapan Algoritme Linear Regression untuk Prediksi Hasil Panen Tanaman Padi," J. Nas. Tek. Elektro dan Teknol. Inf., vol. 8, no. 4, p. 364, 2019.

[13] D. Triseptiawan and R. F. Malik, "Pemantauan Posisi Object Menggunakan Algoritma Multiple Linier Regression," pp. 17–22, 1907.

[14] A. Mei Sarah, B. Kurniadi, and E. Warsini, "Implementasi Metode Regresi Linear Dalam Memprediksi Penyakit Anemia Secara Dini," J. Tek. E-ISSN 2775-0965 (Jurnal Teknol. Komput. dan Sist. Informasi), vol. 3, no. 1, pp. 14–23, 2023.

DOI : https://doi.org/10.52158/jacost.491