Computatio: Journal of Computer Science and Information Systems, 8/1 (2024), 50 - 61

IMPLEMENTASI MACHINE LEARNING UNTUK PREDIKSI HARGA RUMAH MENGGUNAKAN ALGORITMA RANDOM FOREST

Nicholas Hadi[1], Jason Benedict[2]
Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara,
Jln. Letjen S. Parman No. 1, Jakarta, 11440, Indonesia
E-mail: [1]nicholashadi26@gmail.com, [2]jason.benedict2510@gmail.com

Abstrak

Dengan pentingnya peran rumah dalam kehidupan masyarakat, banyak orang yang pasti bertujuan untuk melakukan pembelian atau penjualan rumah. Dengan banyaknya kriteria – kriteria yang dapat mempengaruhi harga rumah, membuat harga rumah sangat susah untuk di prediksi. Harga rumah tersebut tentu saja dapat diprediksi dengan menggunakan 3 algoritma Machine Learning yaitu Random Forest, Decision Tree, dan Polynomial Regression. Manfaat dalam penelitian ini adalah untuk mengetahui kriteria yang paling mempengaruhi harga rumah, dan memperlihatkan hasil akurasi dari setiap algoritma yang digunakan serta menemukan algoritma prediksi terbaik dari 3 algoritma tersebut. Penelitian ini dilakukan pada dataset harga rumah di King County, USA yang bersumber dari situs Kaggle. Dalam hasil pengujian korelasi dari 13 variabel data yang digunakan, ditemukan bahwa variabel luas rumah, grade, dan luas atas rumah mempunyai nilai pengaruh besar terhadap harga rumah. Hasil pengujian 3 algoritma tersebut dievaluasi dengan nilai R^2 dan RMSE. Algoritma Random Forest dinyatakan menghasilkan prediksi terbaik dibandingkan 2 algoritma tersebut, dengan tingkat akurasi sebesar 86,54% dan nilai RMSE sebesar 144913.73.

Kata Kunci—*machine learning, random forest, decision tree, polynomial regression, harga rumah*

This paper describes an algorithm for loan prediction using Polynomial Regression to calculate the amount of debt that can be borrowed by debtors. The advantage of using this algorithm is that the degree selection can be adjusted to the shape of the data distribution so as to get maximum accuracy. We use python programming language along with scikit-learn, numpy, and pandas libraries for prediction and data transformation as well as seaborn and matplotlip for visualization.

Keywords: *machine learning, random forest, decision tree, polynomial regression, house price*

1. PENDAHULUAN

1.1 Latar Belakang

Rumah merupakan salah satu kebutuhan primer manusia. Rata-rata semua manusia melakukan aktifitas sehari-hari mereka di dalam rumah, selain itu juga rumah dapat dijadikan sebagai tempat perlindungan dari gangguan cuaca dan mahluk hidup yang berbahaya. Rumah juga dapat dijadikan alat investasi dengan harga yang dapat berubah dalam waktu tertentu dan keadaan tertentu. Maka dengan itu, banyak orang yang ingin membeli atau melakukan suatu bisnis penjualan rumah. Akan tetapi, harga rumah sangatlah susah untuk di prediksi. Banyak faktor-

faktor kriteria yang dapat mempengaruhi harga rumah tersebut, hal seperti lokasi, fasilitas, banyaknya lantai, besar luas tanah dan rumah, dan masih banyak kriteria yang lainnya. Maka itu, penelitian ini dilakukan untuk melakukan sebuah sistem prediksi harga rumah bedasarkan kriteria tertentu menggunakan *Machine Learning*.

Algoritma yang digunakan dalam penelitian ini adalah algoritma *Random Forest.* Penelitian ini menggunakan algoritma *Random Forest* karena algoritma ini merupakan sebuah metode *ensemble* yang metode ini dapat mendapatkan hasil nilai akurasi yang lebih besar daripada algoritma prediksi lainnya. Pada proses penelitian, akan dibandingkan hasil prediksi Algoritma *Random Forest* dengan 2 algoritma lainya yaitu *Decision Tree* dan *Polynomial Regression.*

1.2 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut :

1. Mendapatkan hasil nilai akurasi menggunakan algoritma *Random Forest* dalam memprediksi harga rumah.
2. Mendapatkan hasil perbandingan antara 3 algoritma yaitu algoritma *Random Forest, Regression, Decision Tree*, dan *Polynomial Regression.*
3. Mengetahui seberapa besar faktor setiap kriteria rumah dalam mempengaruhi harga rumah tersebut.
4. Hasil *web application* yang dapat memprediksi harga rumah.

1.3 Manfaat Penelitian

Manfaat dari hasil penelitian ini adalah untuk dapat membantu masyarakat ataupun pihak industri penjualan rumah dapat secara mudah untuk mengetahui nilai harga estimasi rumah dengan kriteria – kriteria tertentu.

1.4 Penelitian Relevan

Sebelum proses penelitian ini, ditemukan beberapa penelitian serupa yang memakai algoritma *Random Forest*. Penelitian tersebut dapat dijadikan alat bantu dalam menerapkan algoritma *Random Forest* dalam penelitian ini.

Penelitian serupa yang pertama ini dilakukan oleh Aji dan Betha (2018) dengan judul "*Random Forest Algorithm for Prediction of Precipitation*" [1]. Dalam penelitian tersebut, algoritma *Random Forest* digunakan untuk memprediksi curah hujan. Dataset dari penelitian ini diambil di https://www1.ncdc.noaa.gov/pub/orders/ pada bulan Juli 2017 yang terdiri dari 2188 data dengan 16 atribut. Pembagian data latih dan data test dilakukan dengan metode *K-Fold Cross Validation* dengan K=10. Dalam hasil evaluasi pada penelitian ini didapatkan hasil akurasi sebesar 71,6% dengan data yang sudah diproses 10-Fold *Cross Validation* dan dibandingkan dengan data tanpa proses *Fold Cross Validation* menghasilkan akurasi yang sangat tinggi sebesar 99,4%.

Penelitian serupa selanjutnya ini dilakukan oleh Soumi dan Chandan (2020) dengan judul *"A Predictive Analysis Model of Customer Purchase Behavior using Modified Random Forest Algorithm in Cloud Environment*" [2]. Algoritma *Random Forest* digunakan untuk memprediksi pembelian yang dilakukan oleh pelanggan untuk kedepan nanti dalam bentuk *cloud.* Dataset yang dipakai dalam penelitian adalah berupa informasi tentang pelanggan dari Avazu dan riwayat pembeliannya. Dataset tersebut diambil melalui website kaggle. Data tersebut dibagi menjadi 3000 data latih dan 700 data test untuk diproses nanti. Hasil akurasi yang didapatkan pada penelitian ini sebesar 87,02% dengan parameter *Random Forest* 100 pohon. Penelitian ini juga menyatakan bahwa hasil akurasi yang didapakan dari algoritma *Random Forest* lebih besar daripada algoritma *Linear Regression.*

Kemudian Penelitian seruoa yang dilakukan oleh Nariswati, Utami, dan Soni (2011) yang berjudul "Penerapan Metode *Random Forest* dalam *Driver Analysis*" [3]. Penelitian tersebut mencoba anlisis hasil implementasi metode *Random Forest* dalam metode yang bernama "*Driver Analysis"* yang dilakukan untuk memahami pengaruh peubah penjelas terhadap peubah respons sehingga dapat diketahui prioritas setiap peubah penjelas dalam menggerakkan peubah respons. Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari sebuah perusahaan riset pemasaran di Indonesia. Data tersebut terdiri atas sejumlah merek yang berbeda, dimana merek-merek tersebut merupakan jenis produk yang sama, yaitu produk Z. Banyaknya amatan dalam data adalah 1200 amatan. Pada hasil penerapan penelitian tersebut, dihasilkan *driver analysis* yang stabil jika ukuran pohon *random forest* lebih dari 500.

Penelitian serupa selanjutnya dilakukan oleh Widya, Ilham, Muhamad, dan Tri (2021) yang berjudul "Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi *Random Forest*" [4]. Pada penelitian tersebut, algoritma *Random Forest* digunakan untuk memprediksi penyakit diabetes pada tahap awal. Hasil akurasi *Random Forest* tersebut juga nanti di bandingkan oleh metode *SVM* dan *Naïve Bayes*. Data yang digunakan diambil melalui aplikasi bernama WEKA yang dapat mencakup kumpulan berbagai metode pembelajaran mesin untuk klasifikasi data, pengelompokan, regresi, visualisasi, dll. Data yang digunakan terdiri dari 520 jumlah data dan 17 atribut. Pada hasil penelitian tersebut, algoritma *Random Forest* mendapatkan hasil skor akurasi tertinggi dibandingkan metode *SVM* dan *Naïve Bayes.*

Penelitian serupa yang terakhir dilakukan oleh Yogo, Septiadi, dan Anna (2019) dengan judul "Analisis Perbandingan Kinerja *CART* Konvensional *Bagging* dan *Random Forest* pada Klasifikasi Objek: Hasil dari Dua Simulasi" [5]. Tujuan dari penelitian ini adalah untuk membandingkan kinerja dari metode *CART, Bagging,* dan *Random Forest.* Tiga metode tersebut akan memproses dua data simulasi dengan variabel independen yang bertipe non biner dan juga yang bertipe biner. Setelah hasil perhitungan akurasi ketiga metode tersebut, dapat disimpulkan bahwa ketiga metode pada variabel independen bertipe non biner menghasilkan kinerja yang lebih baik dibandingkan variabel independen bertipe biner. Metode Random Forest juga menghasilkan kinerja paling baik dibandingkan CART dan Bagging pada saat variabel independen yang digunakan berisi variabel non biner untuk setiap titik korelasi.

2. METODE PENELITIAN

2.1 Data

Data yang akan digunakan pada penelitian ini adalah data harga rumah di *King County, USA* pada tahun 2014 sampai 2015. Data ini diambil melalui situs kaggle tentang *House Sales in King County, USA*. (https://www.kaggle.com/harlfoxem/housesalesprediction). Data harga rumah yang sudah diambil dan akan digunakan mempunyai total data sebanyak 21613 data, dengan menggunakan sebanyak 13 dari 21 variabel fitur yang dapat mempengaruhi harga rumah tersebut yang akan dijelaskan dalam Tabel 1.

Tabel 1. Penjelasan Variabel Dataset

Nama Variabel	Penjelasan	*DType*
price	Harga rumah (variabel target)	float64
bedrooms	Banyak kamar tidur	int64
bathrooms	Banyak kamar mandi	float64
sqft_living	Luas rumah (ft^2)	int64
sqft_lot	Luas tanah (ft^2)	int64
floors	Banyak tingkat lantai rumah	float64

Nama Variabel	Penjelasan	*DType*
waterfront	Adanya waterfront di daerah rumah tersebut	int64
view	Nilai pemandangan rumah (1-4)	int64
grade	Kelas Rumah (1-13)	int64
sqft_above	Luas bagian atas rumah (ft^2)	int64
sqft_basement	Luas basement rumah (ft^2)	int64
lat	Latitude	float64
long	Longtitude	float64

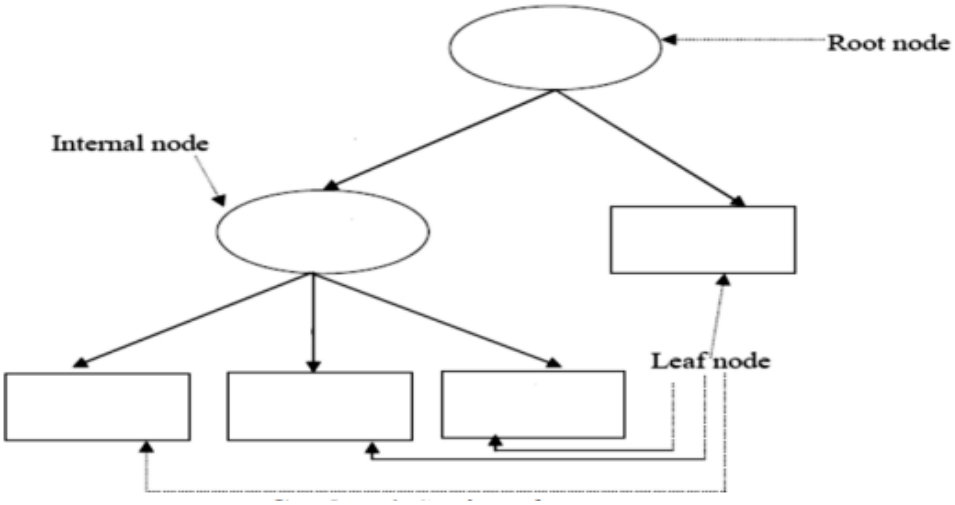
Sebelum data ini digunakan untuk pembuatan model prediksi, data tersebut perlu dilakukan pemrosesan awal data (*data pre-processing*). *Data pre-processing* adalah proses awal data yang dapat membantu model yang digunakan agar dapat menghasilkan nilai output yang lebih baik. *Data pre-processing* yang dilakukan adalah melakukan *cleaning* pada data seperti memperbaiki atau menghapus data yang rusak atau data yang tidak relevan. Perlu juga di deklarasi variabel dependen (Y) yaitu variabel *price* dan varibel independen (X) yaitu 12 variabel lainnya. Data ini dilakukan *spliting*. 70% data tersebut menjadi data *training* dan 30 % sisa data menjadi data *testing*.

2.2 Algoritma

Pada penelitian ini, algoritma yang akan digunakan untuk memprediksi harga rumah adalah algoritma *Random Forest, Decision Tree,* dan *Polynomial Regression*.

2.2.1 *Decision Tree*

Decision Tree merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Sebuah *decision tree* adalah sebuah struktur yang dapat digunakan untuk membagi kumpulan data yang besar menjadi himpunan-himpunan record yang lebih kecil dengan menerapkan serangkaian aturan keputusan. Pada *decision tree* setiap simpul daun menandai label kelas. Simpul yang bukan simpul akhir terdiri dari akar dan simpul internal yang terdiri dari kondisi tes atribut pada sebagian record yang mempunyai karakteristik yang berbeda. Simpul akar dan simpul internal ditandai dengan bentuk oval dan simpul daun ditandai dengan bentuk segi empat [6]. Struktur *decision tree* dapat dilihat pada Gambar 1.



Gambar 1. Struktur *Decision Tree*

2.2.2 *Random Forest*

Algoritma Random Forest adalah pengembangann dari metode *Classification* and *Regression Tree* (CART), yaitu dengan menerapkan metode *bootstrap aggregating* (bagging) dan *random feature selection* [7]. 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 diatasnya) dan dilakukan pemilihan secara acak.

Kemudian hasil klasifikasi dari setiap pohon diakumulasikan dan dipilh hasil klasifikasi yang paling banyak muncul [8]. Metode ini terdiri dari *root node*, *internal node*, dan *leaf node. Root node* merupakan simpul yang terletak paling atas, atau biasa disebut sebagai akar dari pohon keputusan. Internal node adalah simpul percabangan, dimana node ini mempunyai output minimal dua dan hanya ada satu input. Sedangkan leaf node atau terminal node merupakan simpul terakhir yang hanya memiliki satu input dan tidak mempunyai output. Pohon keputusan dimulai dengan cara menghitung nilai entropy sebagai penentu tingkat ketidakmurnian atribut dan nilai information gain. Untuk menghitung nilai entropy digunakan rumus seperti pada persamaan (1), sedangkan nilai *information gain* menggunakan persamaan (2)[9].

$$Entropy(Y) = -\sum_i p(c|Y) \log 2p(c|Y), \quad (1)$$

Keterangan :
Y = Himpunan kasus
P(c|Y) = Proporsi nilai Y terhadap kelas c.

$$Information\ Gain(Y, a) = Entropy(Y) - \sum_{v \epsilon Values(a)} \frac{|Y_v|}{|Y_a|} Entropy(Y_v), \quad (2)$$

Keterangan :
Values(a) = Nilai yang mungkin dalam himpunan kasus a.
Y_v = Subkelas dari Y dengan kelas v yang berhubungan dengan kelas a.
Y_a = Semua nilai yang sesuai dengan a.

2.2.3 *Polynomial Regression*

Polynomial Regression merupakan model regresi linier yang dibentuk dengan menjumlahkan pengaruh masing-masing variabel prediktor (X) yang dipangkatkan meningkat sampai orde ke-k. Secara umum, model *Polynomial Regression* ditulis dalam bentuk persamaan (3). Model *Polynomial Regression* untuk variabel prediktor berganda dapat berbentuk dalam persamaan (4) [10].

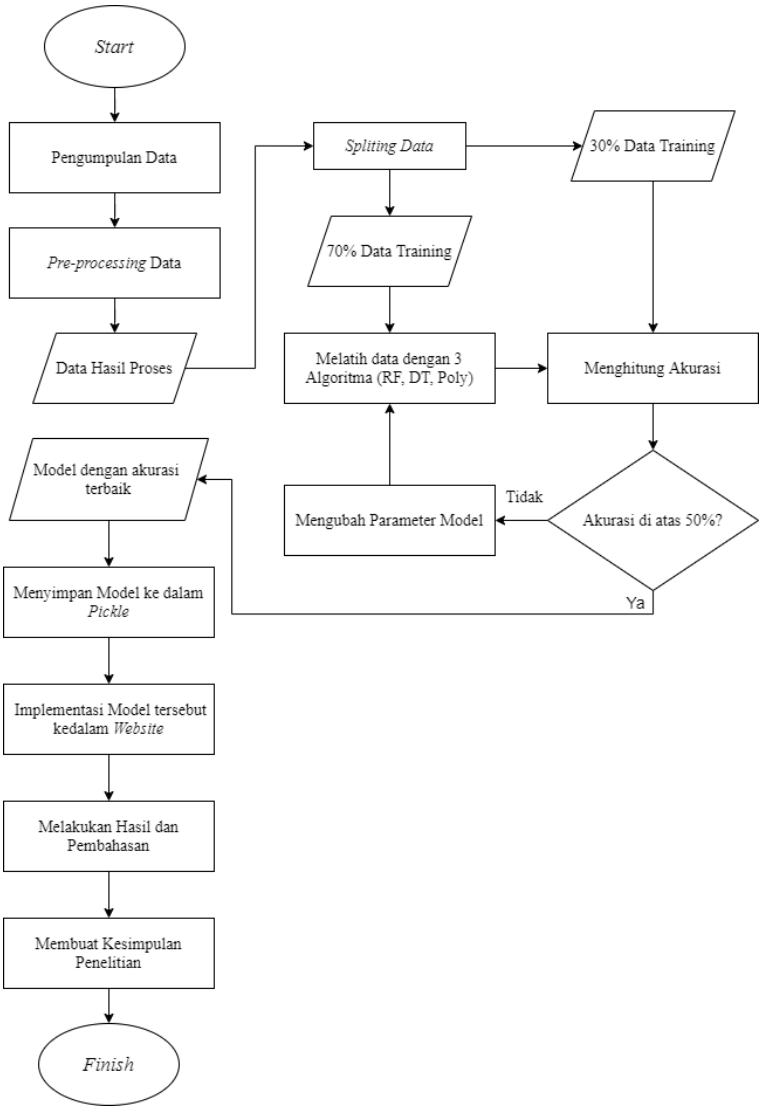
$$Y = b_0 + b_1 X + b_2 X^2 + ... + b_k X^k + \varepsilon, \quad (3)$$

$$Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k + \varepsilon, \quad (4)$$

Keterangan :
Y = Variabel respons
b_0 = Intersep
$b_1, b_2, ..., b_k$ = Koefisien-koefisien regresi
X = Variabel prediktor
ε = Faktor pengganggu yang tidak dapat dijelaskan oleh model regresi

2.3 Rancangan Penelitian

Tahapan atau Rancangan yang dilakukan dalam penelitian ini digambarkan dalam bentuk *Flowchart* yang dapat dilihat pada Gambar 2.



Gambar 2. Flowchart Rancangan Penelitian

2.4 Metode Evaluasi

Metode evaluasi yang akan dipakai dalam penelitian ini adalah dengan menghitung nilai R^2 (Koefisien Determinasi), nilai RMSE (*Root Mean Square Error*), dan *Scatter Plot*. Nilai R2 menunjukan persentase variabel tak bebas dapat dijelaskan oleh variabel bebas. Nilai koefisien determinasi adalah $0 < R^2 < 1$. Semakin tinggi nilai R^2 maka semakin baik model karena semakin besar keragaman peubah dependen yang dapat dijelaskan oleh peubah independen. Perhitungan R^2 dapat dilakukan dengan persamaan (5) [11].

$$R^2 = 1 - \frac{SS\ error}{SS\ total} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (5)$$

Keterangan :

y_i = Observasi respons ke-i

$\hat{y}_i$ = Ramalan respons ke-i

$\bar{y}$ = rata-rata

RMSE (Root Mean Squared Error) sering digunakan untuk mengevaluasi kinerja prediksi yang digunakan dengan mengukur tingkat akurasi dari hasil prediksi. Jika nilai RMSE rendah

menunjukan bahwa bentuk perubahan nilai yang dihasilkan oleh suatu model prediksi mendekati bentuk nilai aslinya. Dan sebaliknya, jika besar nilai RMSE maka keakuratan yang dihasilkan jauh dari bentuk nilai aslinya. Perhitungan RMSE dapat dilakukan dengan persamaan (6) [11].

$$RMSE = \frac{\sum_i^n = \sqrt{(y_i - \hat{y}_i)^2}}{n} \quad (6)$$

Metode evaluasi *Scatter Plot* digunakan untuk menggabungkan nilai hasil yang diprediksi sebagai *plot* titik yang di *scatter* dan nilai uji sebagai *plot* garis, dalam satu *figure.* Visual *figure* tersebut dapat memperlihatkan besar *outlier* nilai yang sudah di prediksi dengan nilai uji.

2.4 *Web Framework*

Semua proses penelitian ini dilakukan dengan menggunakan kode *Python*, maka *web framework* yang akan digunakan dalam penelitian ini adalah *framework* yang bernama *streamlit*. *Streamlit* adalah *Web Application Framework* untuk memudahkan *developer* dalam pengembangan dan pembuatan aplikasi web yang fokus di bidang *Machine Learning* dan *Data Science.* Tidak hanya itu, *Streamlit* dapat secara mudah melakukan *hosting* aplikasi web yang sudah dibangun. *Streamlit* tersedia dalam bentuk *Library* dan dapat di-install melalui *Python*.

3. HASIL DAN PEMBAHASAN

3.1 Pengujian Korelasi

Pengujian nilai variabel korelasi dalam penelitian ini bermanfaat untuk mengetahui seberapa besar pengaruh setiap 12 variabel independent terhadap variabel dependen yaitu varibel *price.* Semakin besar nilai korelasi maka semakin besar pengaruh variabel tersebut terhadap hasil prediksi harga rumah.

	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	grade	sqft_above	sqft_basement	lat	long
price	1.00	0.31	0.53	0.70	0.09	0.26	0.27	0.40	0.67	0.61	0.32	0.31	0.02
bedrooms	0.31	1.00	0.52	0.58	0.03	0.18	-0.01	0.08	0.36	0.48	0.30	-0.01	0.13
bathrooms	0.53	0.52	1.00	0.75	0.09	0.50	0.06	0.19	0.66	0.69	0.28	0.02	0.22
sqft_living	0.70	0.58	0.75	1.00	0.17	0.35	0.10	0.28	0.76	0.88	0.44	0.05	0.24
sqft_lot	0.09	0.03	0.09	0.17	1.00	-0.01	0.02	0.07	0.11	0.18	0.02	-0.09	0.23
floors	0.26	0.18	0.50	0.35	-0.01	1.00	0.02	0.03	0.46	0.52	-0.25	0.05	0.13
waterfront	0.27	-0.01	0.06	0.10	0.02	0.02	1.00	0.40	0.08	0.07	0.08	-0.01	-0.04
view	0.40	0.08	0.19	0.28	0.07	0.03	0.40	1.00	0.25	0.17	0.28	0.01	-0.08
grade	0.67	0.36	0.66	0.76	0.11	0.46	0.08	0.25	1.00	0.76	0.17	0.11	0.20
sqft_above	0.61	0.48	0.69	0.88	0.18	0.52	0.07	0.17	0.76	1.00	-0.05	-0.00	0.34
sqft_basement	0.32	0.30	0.28	0.44	0.02	-0.25	0.08	0.28	0.17	-0.05	1.00	0.11	-0.14
lat	0.31	-0.01	0.02	0.05	-0.09	0.05	-0.01	0.01	0.11	-0.00	0.11	1.00	-0.14
long	0.02	0.13	0.22	0.24	0.23	0.13	-0.04	-0.08	0.20	0.34	-0.14	-0.14	1.00

Gambar 3 Nilai Korelasi dari Setiap Variabel

Pada Gambar 3 dapat dilihat variabel dengan nilai korelasi yang tinggi terhadap variabel dependen (*price*) adalah variabel luas rumah (*sqft_living*) dengan nilai korelasi sebesar 0,70, variabel kelas rumah (*grade*) dengan nilai korelasi sebesar 0,67, dan luas atas rumah (sqft_above) dengan nilai korelasi sebesar 0,61. Hal ini menunjukan bahwa 3 variabel tersebut memiliki dampak yang lebih besar dalam mempengaruhi harga rumah. Adapun juga variabel yang memiliki

nilai korelasi yang kecil yaitu variabel luas tanah (*sqft_lot*) dengan nilai korelasi sebesar 0,09, dan variabel longtitude (*long*) dengan nilai korelasi sebesar 0.02.

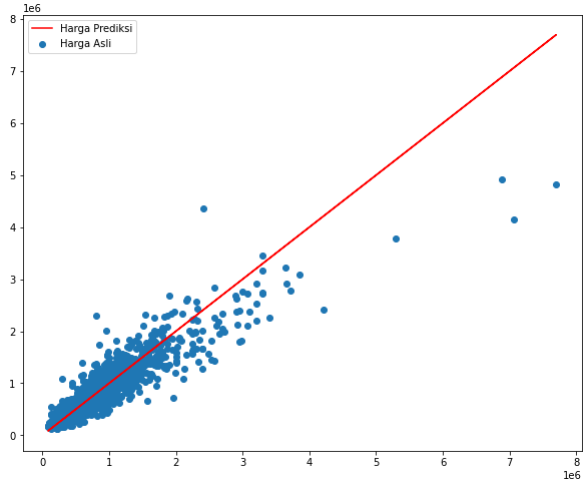
3.2 Hasil Metode *Random Forest*

Dari data yang sudah dikumpul sebanyak 21613 jumlah data, dan sudah melakukan pembagian data berasio 7:3, dimana 70% adalah data latih dan 30% adalah data uji. Mengunakan algoritma *Random Forest* harus diawali dengan pemilihan parameter yaitu jumlah pohon (N*tree*) yang akan digunakan pada proses pelatihan model tersebut. Jumlah pohon (N*tree*) yang akan dipilih dalam penelitian ini yaitu sebanyak 10, 20, dan 30. Selanjutnya melakukan proses perhitungan nilai evaluasi yaitu tingkat akurasi (R^2) dan nilai RMSE (*Root Mean Square Error*) dengan menggunakan data uji.

Tabel 2 Hasil Evaluasi setiap Ntree yang dipilih

Jumlah Pohon (N*Tree)*	R^2 (%)	RMSE
10	85,05	152722.58
20	86,54	144913.73
30	86,34	145977.42

Pada Tabel 2 dapat dilihat menggunakan parameter jumlah pohon (N*tree*) hanya sebanyak 10 sudah memberikan hasil yang memuaskan dengan R^2 sebesar 0, 8505 atau 85,05%, dan dengan RMSE sebesar 152722.58. Hasil model prediksi *Random Forest* yang terbaik menggunakan parameter jumlah pohon (N*tree*) sebanyak 20, dengan menghasilkan tingkat akurasi prediksi yang paling besar yaitu 86,54% dan nilai RMSE yang paling terkecil sebesar 145913,42. *Scatter Plot* pada Gambar 4 dapat dilihat outlier antara data harga uji dengan data harga yang di prediksi menggunakan model *Random Forest* dengan parameter terbaik. Dalam *Scatter Plot* ini, dapat dilihat jarak *plot* titik nilai harga prediksi tidak sangat jauh dengan plot garis nilai harga asli, ini menyatakan bahwa model *Random Forest* ini mempunyai hasil yang akurat Hasil model terbaik tersebut akan nanti dibandingkan dengan algoritma *Decision Tree* dan *Polynomial Regression.*

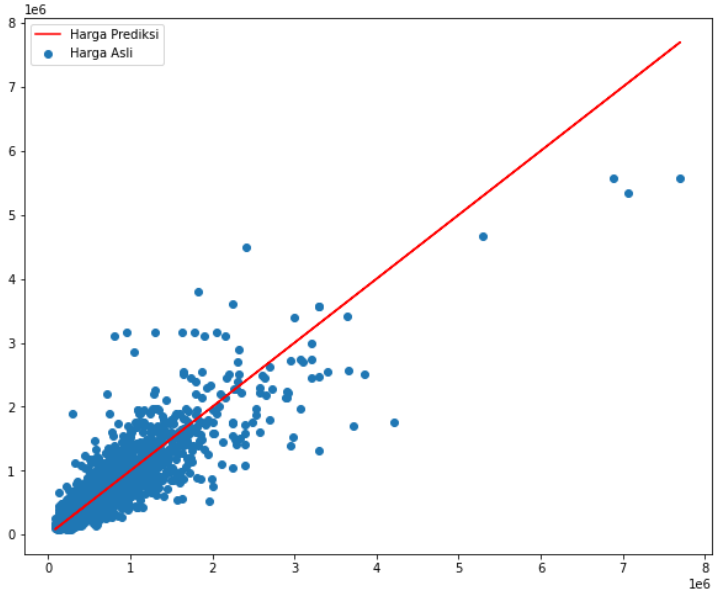


Gambar 4 Scatter Plot Model Random Forest dengan Ntree=20

3.2 Perbandingan Hasil *Random Forest* dengan *Decision Tree* dan *Polynomial Regression*

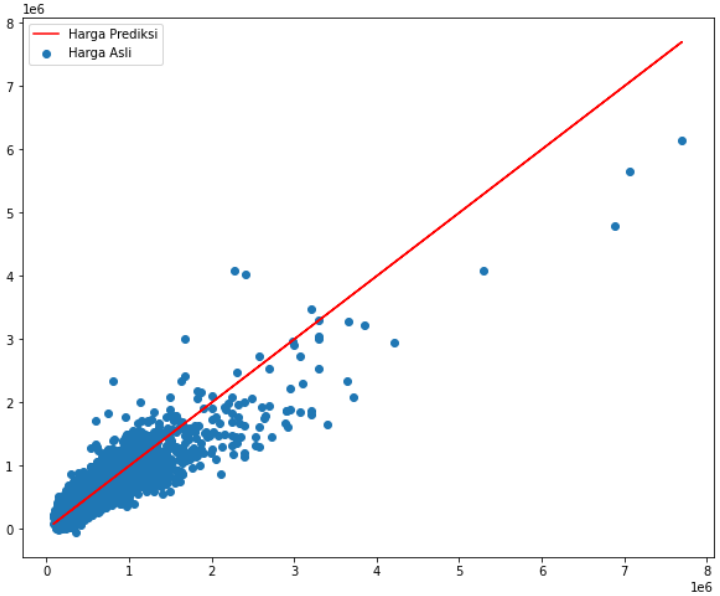
Hasil prediksi model *Random Forest* yang sudah didapatkan akan di lakukan perbandingan dengan 2 algoritma lainnya. Algoritma yang pertama yang digunakan adalah algoritma *Decision Tree*. Dengan menggunakan pembagian data latih dan data uji yang sama, model prediksi *Decision Tree* menghasilkan tingkat akurasi (R^2) sebesar 0,7639 atau 76,39%, serta menghasilkan

nilai RMSE sebesar 191920,88. *Scatter Plot* model prediksi *Decision Tree* untuk melihat besar outlier data harga prediksi dengan data harga uji dapat dilihat pada Gambar 5. Dilihat dari *Scatter Plot* ini, *plot* titik harga prediksi mempunyai *outlier* dengan garis harga asli lebih banyak dan lebih jauh dibandingkan *Scatter Plot* Gambar 4.



Gambar 5 Scatter Plot Model Decision Tree

Model algoritma selanjutnya adalah model *Polynomial Regression*. Dengan menggunakan data latih dan data uji yang sama, model *Polynomial Regression* dengan menggunakan parameter *degree* sebesar 2 mendapatkan tingkat akurasi (R^2) sebesar 0,7813 atau 78,13%, serta menghasilkan nilai MRSE sebesar 184708,77. *Scatter Plot* model prediksi *Polynomial Regression* untuk melihat besar outlier data harga prediksi dengan data harga uji dapat dilihat pada Gambar 6. Sama seperti *Scatter Plot* pada Gambar 5, *Scatter Plot Polynomial Regression* ini mempunyai nilai outlier yang lebih banyak dibandingkan *Scatter Plot Random Forest* pada Gambar 4.



Gambar 6 Scatter Plot Model Polynomial Regression

Tabel 3 Hasil Perbandingan Model Prediksi Harga Rumah

Model	R^2 (%)	RMSE
Random Forest	86,54	144913.73
Decision Tree	76,39	191920,88
Polynomial Regression	78,13	184708,77

Pada Tabel 3 dapat dinyatakan bahwa dalam penelitian prediksi harga rumah ini, model algoritma yang menghasilkan nilai prediksi paling akurat adalah algoritma *Random Forest.* Model *Random Forest* ini akan disimpan dengan *pickle* agar dapat dibaca pada *framework web app streamlit*.

3.2 Hasil *User Interface Web Application*

Dengan model *Machine Learning* telah disimpan dengan *pickle* lalu dapat dimasukan ke dalam *framework* yang bernama *streamlit*, dapat dibangun sebuah aplikasi web yang dapat digunakan oleh *user* untuk melakukan prediksi harga rumah. Aplikasi web tersebut sudah di-*host* dengan *streamlit* yang dapat dibuka pada link: https://share.streamlit.io/nicholashd/thebojongershousepredictionapp/main.py. Tampilan halaman web dapat dilihat pada Gambar 7.

Gambar 7 Tampilan halaman aplikasi web

Dalam aplikasi web ini, *user* dapat langsung saja memasukan input kriteria rumah di *form* yang berada di sebelah kiri web tersebut. Setelah semua kriteria sudah di *input*, *user* dapat menekan tombol "*Predict Now"* untuk aplikasi web tersebut mengeluarkan hasil *output* harga dari kriteria yang sudah di input oleh *user.* Tampilan hasil *ouput* harga rumah dapat dilihat pada Gambar 8.

Gambar 8 Tampilan Hasil Output

Untuk *user* yang ingin melakukan prediksi ulang, maka user cukup mengubah input kriteria dalam *form* tersebut. Jika *user* ingin melihat kode Python untuk membangun web aplikasi serta model *Machine Learning* yang digunakan, *user* dapat membuka *github repository owner* dengan meng-klik *hyperlink* "*github repository"* yang bewarna biru.

4. KESIMPULAN

Bedasarkan hasil pengujian yang sudah dilakukan dalam penelitian, dapat di ambil kesimpulan bahwa :

1. Setelah melakukan pengujian korelasi, variabel yang mempunyai pengaruh besar terhadap variabel harga rumah adalah variabel luas rumah, grade, dan luas atas rumah.
2. Dari hasil pengujian 3 model *Random Forest* dengan parameter jumlah pohon (N*tree*) yang berbeda*,* kami menemukan bahwa model *Random Forest* dengan jumlah pohon (N*tree*) sebanyak 20 menghasilkan hasil prediksi terbaik dengan tingkat akurasi (R^2) sebesar 86,54% dan nilai RMSE sebesar 144913.73 yang dibandingkan model *Random Forest* lainnya dengan jumlah pohon sebanyak 10 dan 30.
3. Dari hasil perbandingan antara algoritma *Random Forest* dengan *Decision Tree* dan *Polynomial Regression,* dapat diambil bahwa algoritma *Random Forest* dengan 20 jumlah pohon adalah model prediksi yang menghasilkan prediksi terbaik dibandingkan dengan hasil model *Decision Tree* dengan tingkat akurasi sebesar 76,39% dan model *Polynomial Regression* dengan tingkat akurasi sebesar 78,13%.

Untuk penelitian dan pengujian kedepan, hal yang dapat diperhatikan agar mendapatkan hasil yang lebih baik adalah:

1. Dapat menambahkan variabel fitur lebih banyak dalam dataset yang digunakan.
2. Menambahkan algoritma *Machine Learning* lain untuk melakukan perbandingan hasil yang lebih luas
3. Menambahkan fitur pada aplikasi web *interface*.

DAFTAR PUSTAKA

[1] P. Aji and N. S. Betha, "Random Forest Algorithm for Prediction of Precipitation," *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM),* vol. 1, no. 1, pp. 27-31, 2018.

[2] G. Soumi and B. Chandan, "A Predictive Analysis Model of Customer Purchase Behavior using Modified Random Forest Algorithm in Cloud Environment," *IEEE International Conference for Convergence in Engineering,* pp. 239-244, 2020.

[3] K. D. Nariswari, D. S. Utami and Y. Soni, "Penerapan Metode Random Forest dalam Driver Analysis," *Forum Statistika dan Komputasi,* vol. 16, no. 1, pp. 35-43, 2011.

[4] A. Widya, K. Ilham, B. Muhamad and H. Tri, "Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest," *SISTEMASI: Jurnal Sistem Informasi,* vol. 10, no. 1, pp. 163-171, 2021.

[5] A. J. Yogo, P. Septiadi and C. Anna, "Analisis Perbandingan Kinerja CART Konvensional, Bagging dan Random Forest pada Klasifikasi Objek : Hasil dari Dua Simulasi," *Media Statistika,* vol. 12, no. 1, pp. 1-12, 2019.

[6] M. Ari and A. W. Rika, "Model Data Mining sebagai Prediksi Penyakit Hipertensi Kehamilan dengan Teknik Decision Tree," Scientific Journal of Informatics, vol. 3, no. 1, pp. 20-21, 2016.

[7] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[8] A. K. Fransiska and P. K. Angelina, "Analisis Dan Implementasi Random Forest dan Classification dan Regression Tree (CART) untuk Klasifikasi pada Misuse Intrusion Detection System," p. 7, 2011.

[9] S. N. Yusuf and E. Nova, "Sistem Klasifikasi Variabel Tingkat Penerimaan Konsumen Terhadap Mobil Menggunakan Metode Random Forest," vol. 9, no. 1, p. 6, 2017.

[10] S. M. Julyanti, K. Hanny and H. Djoni, "Pengembangan Model Regresi Polinomial Berganda Pada Kasus Data Pemasaran," vol. 12, no. 2, p. 150, 2012.

[11] S. Andi, A. Septi and G. Aris, "Prediksi Harga Rumah Menggunakan Web Scrapping," *Jurnal Teknik Informatika dan Sistem Informasi,* vol. 8, no. 1, pp. 45-46, 2021.