Laporan Proyek Machine Learning - Grivia Trifosa Iskandar

Domain Proyek

Teknologi memiliki peran besar dalam kehidupan manusia saat ini, termasuk salah satunya yaitu penggunaan Machine Learning untuk membantu manusia dalam menyelesaikan permasalahan komputasi yang rumit. Salah satu contohnya yaitu penggunaan Machine Learning untuk memprediksi harga rumah di Jakarta Selatan.

Menurut Rahayuningtyas, rumah merupakan kebutuhan yang diperlukan bagi manusia sebagai tempat tinggal. Dalam kebutuhan untuk membeli rumah, terdapat beberapa aspek yang dapat dipertimbangkan untuk memberikan harga pada rumah. Dengan adanya penggunaan teknologi dalam memprediksi harga rumah, diharapkan agar dapat menghitung korelasi dari berbagai aspek-aspek pada rumah tersebut, sehingga dapat memberikan informasi mengenai harga rumah yang sesuai dengan keadaan.

Berdasarkan dataset ini, akan dilatih model Machine Learning yang mampu untuk memprediksi harga rumah yang ada di Tebet, Jakarta Selatan. Penulis akan menyelesaikan permasalahan prediksi ini dengan model regressi, dan model akan menghasilkan harga rumah berdasarkan data yang telah dibagi menjadi data *train* dan data *test*.

Business Understanding

Problem Statements

- Apakah model dapat memprediksi harga rumah di daerah Tebet, Jakarta Selatan dengan baik?

Goals

- Mengetahui model yang terbaik untuk memprediksi harga rumah di daerah Tebet, Jakarta Selatan

Solution Statements

- Menggunakan EDA untuk dapat melihat fitur yang berkorelasi dan memiliki pengaruh terhadap harga rumah
- Menggunakan Model Machine Learning yang sesuai, yaitu regresi. Terdapat beberapa model yang akan digunakan untuk melihat model mana yang akan menghasilkan nilai prediksi harga rumah yang terbaik. Berikut model-model yang akan digunakan:
  - *K-Neighbors Regressor*
  - *AdaBoost Regressor*
  - *Random Forest Regressor*

Data Understanding

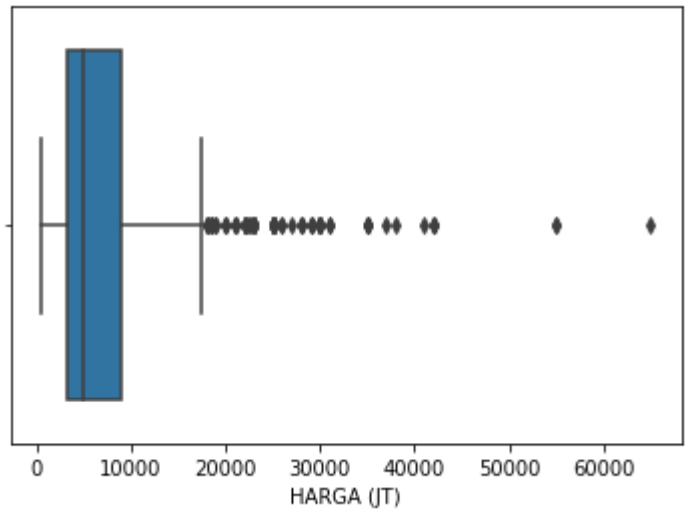
Dataset yang digunakan ini diambil dari platform Kaggle yang dipublikasikan oleh Wisnu Anggara. Dataset ini terdiri dari 2 file *.xlsx* yaitu data untuk rumah yang ada di Jakarta Selatan dan data untuk rumah yang ada di Tebet, salah satu daerah di Jakarta Selatan. Berikut akses link menuju dataset www.kaggle.com/datasets/wisnuanggara/daftar-harga-rumah. Data pada tiap file dataset

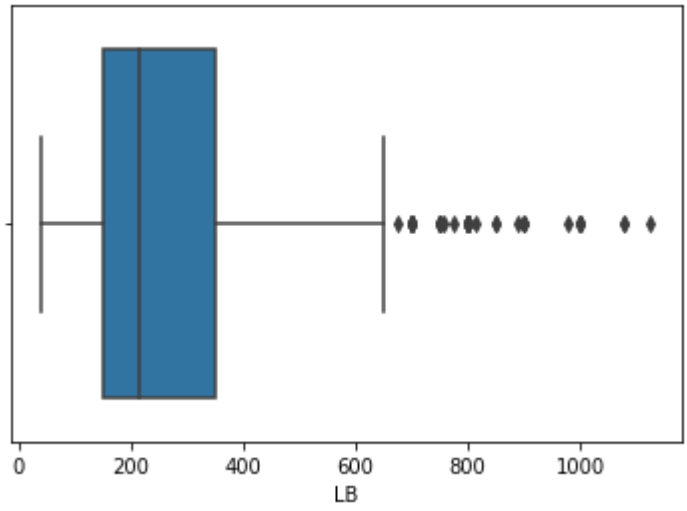
memiliki sekitar 1000 data. Proyek ini menggunakan file dataset untuk rumah yang ada di Tebet, dengan penjelasan fitur sebagai berikut.

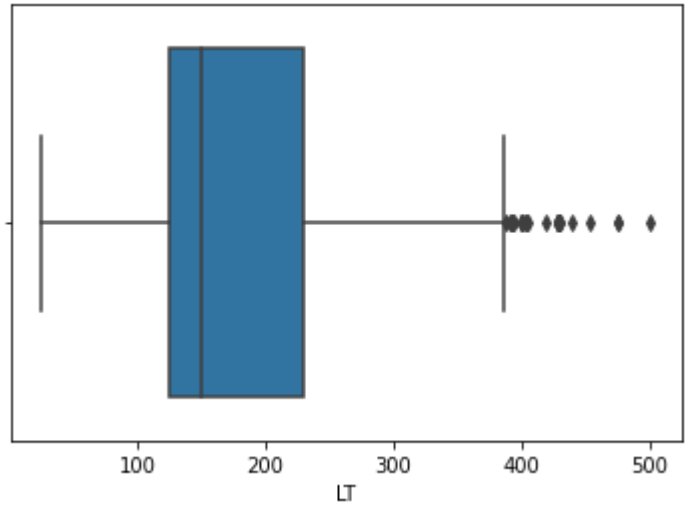
- NO : nomor data.
- NAMA RUMAH : title rumah.
- HARGA : harga dari rumah.
- LB : jumlah luas bangunan.
- LT : jumlah luas tanah.
- KT : jumlah kamar tidur.
- KM : jumlah kamar mandi.
- GRS : jumlah kapasitas mobil dalam garasi.

Selain itu, dilakukan juga *Exploratory Data Analysis* (EDA) yang bertujuan untuk menghilangkan outliers, serta menampilkan korelasi antar data baik data kategorikal maupun data numerik.

Berikut merupakan visualisasi boxplot dari data numerik dari LT, LB, dan Harga (JT).

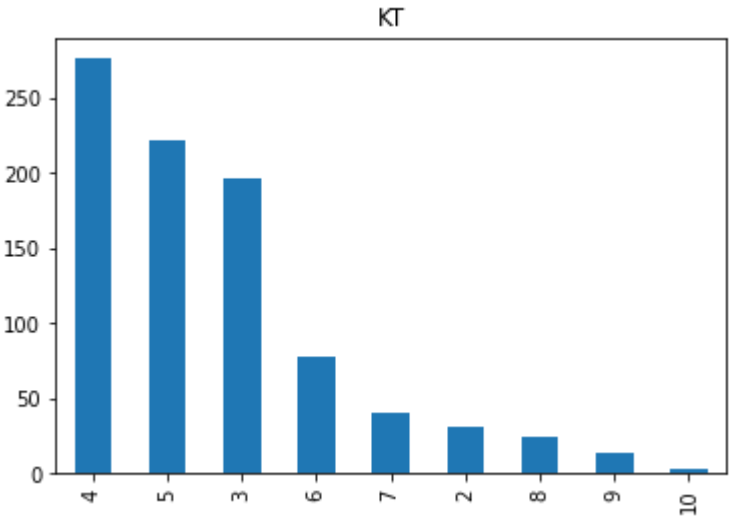




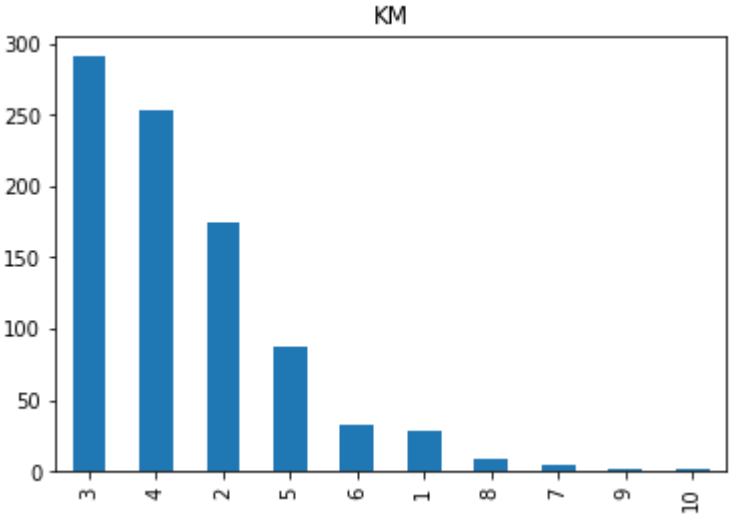


Dapat dilihat dari ketiga gambar tersebut, semua fitur memiliki outliers. Oleh karena itu, digunakan metode *Interquartile Range* (IQR) untuk mengatasi outliers tersebut. Sehingga, nantinya data akan direduksi dan dieliminasi guna mengatasi outliers.

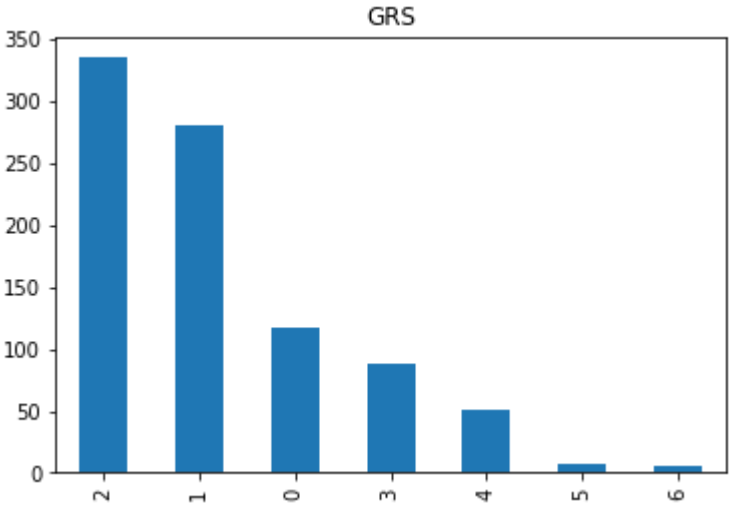
Selanjutnya, dilakukan *univariate analysis* untuk data kategorikal dan data numerik.



Dari gambar diatas dapat disimpulkan bahwa pada fitur KT (Jumlah Kamar Tidur), data terbanyak ditempati oleh 4 kamar tidur dan data tersedikit ditempati oleh 10 kamar tidur.

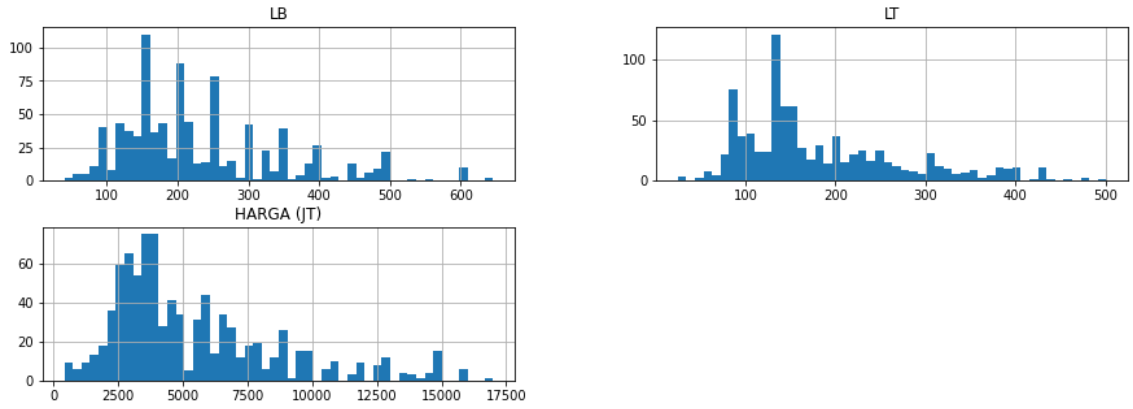


Dari gambar diatas dapat disimpulkan bahwa pada fitur KM (Jumlah Kamar Mandi), data terbanyak ditempati oleh 3 kamar mandi dan data tersedikit ditempati oleh 10 kamar mandi.



Dari gambar diatas dapat disimpulkan bahwa pada fitur GRS (Jumlah Mobil yang Muat Dalam Garasi), data terbanyak ditempati oleh 2 mobil dan data tersedikit ditempati oleh 6 mobil.

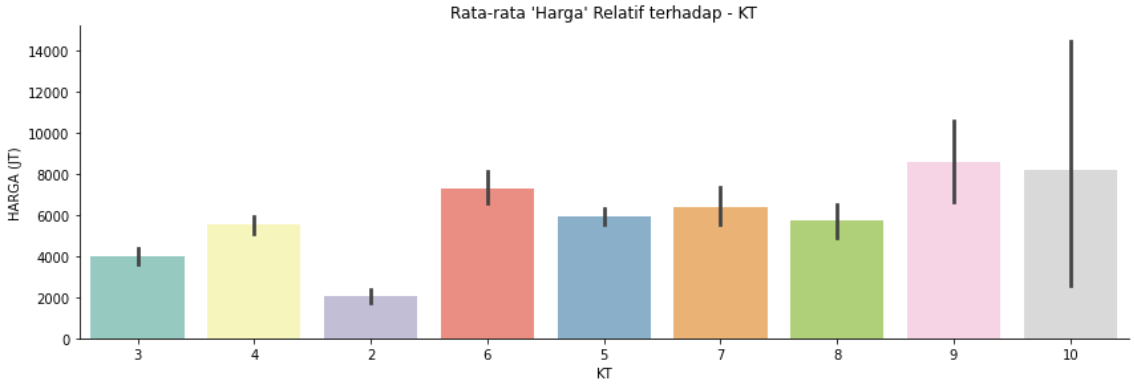
Visualisasi data numerik dilakukan dengan menggunakan plot histogram.

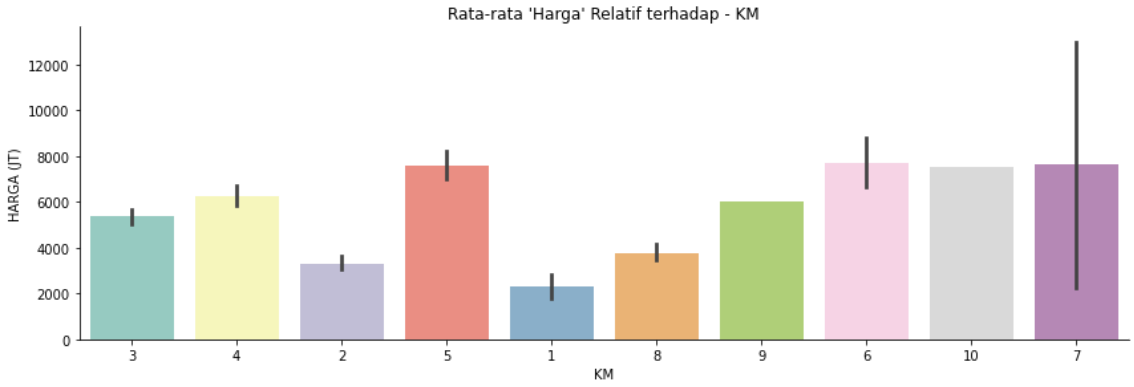


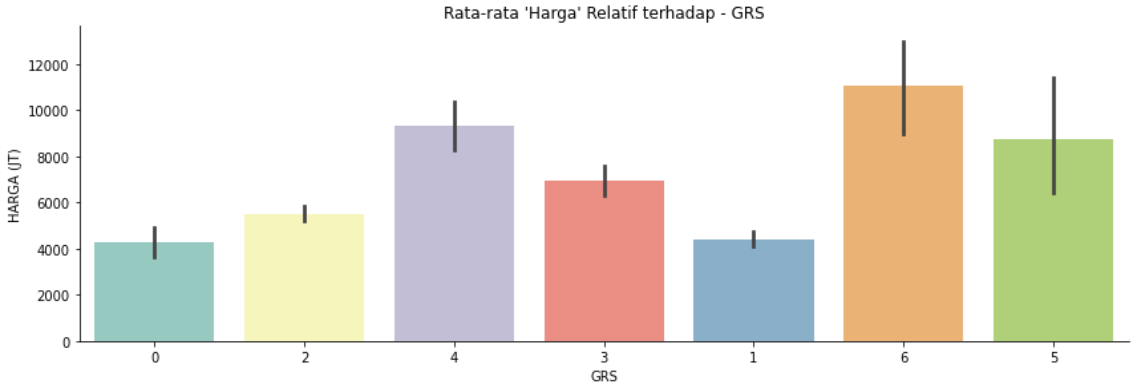
Dari gambar diatas, dapat ditarik kesimpulan, yaitu:

- Pada data "Harga (JT)", data rumah kebanyakan terdapat direntang 2.500.000.000 hingga 5.000.000.000
- Distribusi data miring ke kanan (right skewed) yang dimana akan berdampak pada hasil prediksi model.

Selain itu, terdapat juga EDA *multivariate analysis* untuk data kategorikal dan data numerik.



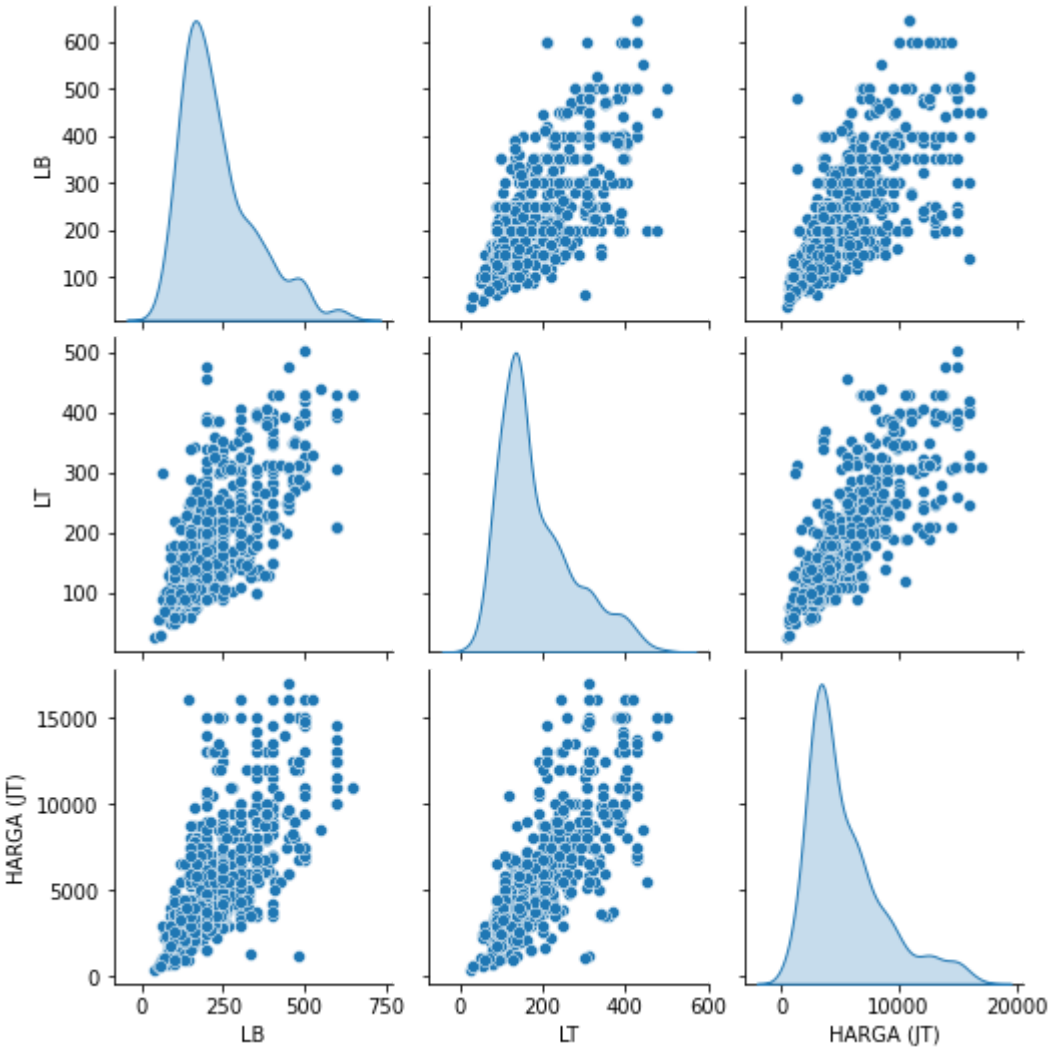




Dari data diatas, dapat disimpulkan:

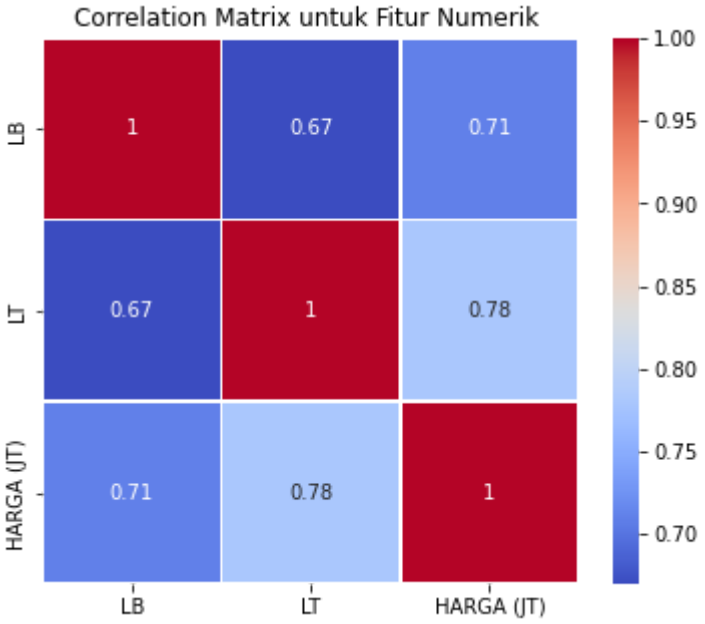
- Data pada KT (Jumlah Kamar Tidur), jumlah kamar tidur 10 memiliki nilai yang rendah, sehingga dapat disimpulkan fitur KT memiliki pengaruh dampak yang kecil terhadap rata-rata harga.
- Data pada KT (Jumlah Kamar Mandi), jumlah kamar mandi 10 memiliki nilai yang mirip dengan nilai yang lain, sehingga dapat disimpulkan fitur KM memiliki pengaruh dampak yang kecil terhadap rata-rata harga.
- Data pada GRS (Jumlah Mobil yang Muat di Garasi), dimana 6 mobil yang muat di garasi memiliki nilai tertinggi dibanding yang lain, sehingga dapat disimpulkan fitur GRS memiliki dampak terhadap rata-rata harga.

Pada data numerik, digunakan pairplot untuk melihat hubungan antara data fitur dan data target.



Dapat disimpulkan berdasarkan gambar diatas, bahwa fitur "LB" dan "LT" memiliki hubungan data yang positif dengan data "Price".

Serta terdapat juga heatmap yang bertujuan untuk memvisualisasikan korelasi antara fitur "LB" dan "LT" dengan data "Price" agar lebih mudah untuk dilihat dan dipahami.



Data Preparation

- Mengatasi outliers dengan menggunakan metode *Interquartile Range* (IQR) yang akan berdampak pada pengurangan data pada dataset.
- Melakukan one hot encoding pada data-data kategorikal dengan menggunakan fungsi *get_dummies* pada library Pandas, dimana data diubah menjadi bilangan biner.
- Membagi data menjadi data train dan data test dengan fungsi *train_test_split*. Pembagian data dilakukan sebanyak 80% untuk data train dan 20% untuk data test.
- Melakukan standarisasi untuk fitur numerik agar menghasilkan nilai standar deviasi sama dengan 1 dan mean sama dengan 0. Standarisasi dilakukan agar memudahkan algoritma dalam melakukan komputasi perhitungan.

Modeling

- Berikut penjelasan beberapa algoritma yang membantu dalam pembuatan model Machine Learning, dimana algoritma yang diambil merupakan algoritma bertipe regresi.
    - **Random Forest**, merupakan salah satu algoritma populer yang digunakan karena kesederhanaannya dan memiliki stabilitas yang baik.
    - **K-Neighbors Regressor**, merupakan salah satu algoritma yang didasari oleh K-Nearest Neighbors. Algoritma ini memiliki kelebihan yaitu dapat melakukan komputasi yang baik pada data yang bersifat non-linear. Namun algoritma ini juga memiliki kelemahan yaitu sensitif terhadap noise seperti missing value atau outliers.
    - **AdaBoost**, merupakan singkatan dari Adaptive Boosting. Algoritma ini bertujuan untuk memberikan bobot lebih pada observasi yang tidak tepat atau disebut weak classification.
- Berikut merupakan tahapan pembuatan model dengan beberapa algoritma yang berbeda.

1. Sebelum membuat model, dilakukan dulu pembuatan DataFrame yang akan diisi dengan hasil MSE data train dan data test pada setiap algoritma.
2. Selanjutnya, dilakukan pembuatan model Random Forest dengan melakukan import library pada sklearn.ensemble yang mengambil fungsi RandomForestRegressor. Setelah itu membuat model dengan diisikan beberapa parameter seperti n_estimators=150, max_depth=16, dan random_state=100.
3. Pada algoritma Boosting, melakukan import library sklearn.ensemble yang mengambil fungsi AdaBoostRegressor. Digunakan beberapa parameter seperti n_estimators=50, learning_rate=0.001, dan random_state=100.
4. Pada tahapan ini, dilakukan import library sklearn.neighbors yang mengambil fungsi KNeighborsRegressor. Pada algoritma K-Neighbors Regressor, digunakan parameter n_neighbors=13.

*Catatan: pada nilai yang terdapat pada tiap parameter diisi dengan angka acak dimana dilakukan *trial dan error* beberapa kali hingga mendapatkan nilai MSE yang terkecil dari hasil tersebut.*

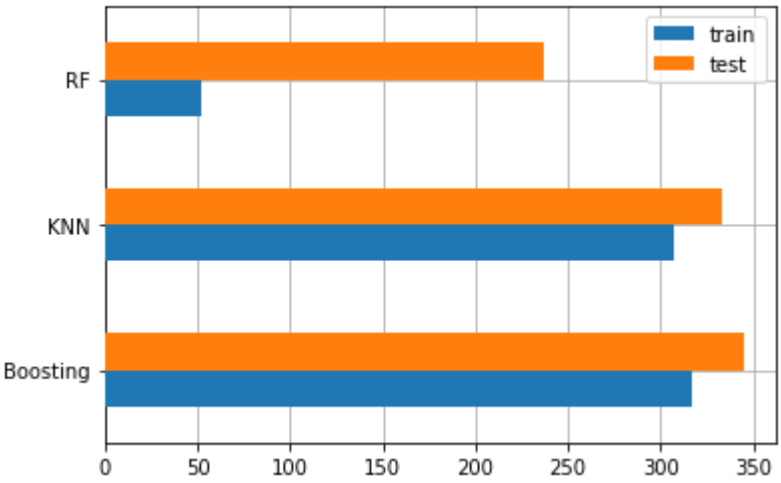
Evaluation

Pada tahap evaluasi, digunakan Mean Squared Error (MSE) yang berfungsi untuk menghitung rata-rata jumlah selisih kuadrat rata-rata nilai sebenarnya dengan nilai prediksi, serta memiliki formula sebagai berikut

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE = mean squared error n = jumlah dataset Yi = nilai sebenarnya Ŷi = nilai prediksi

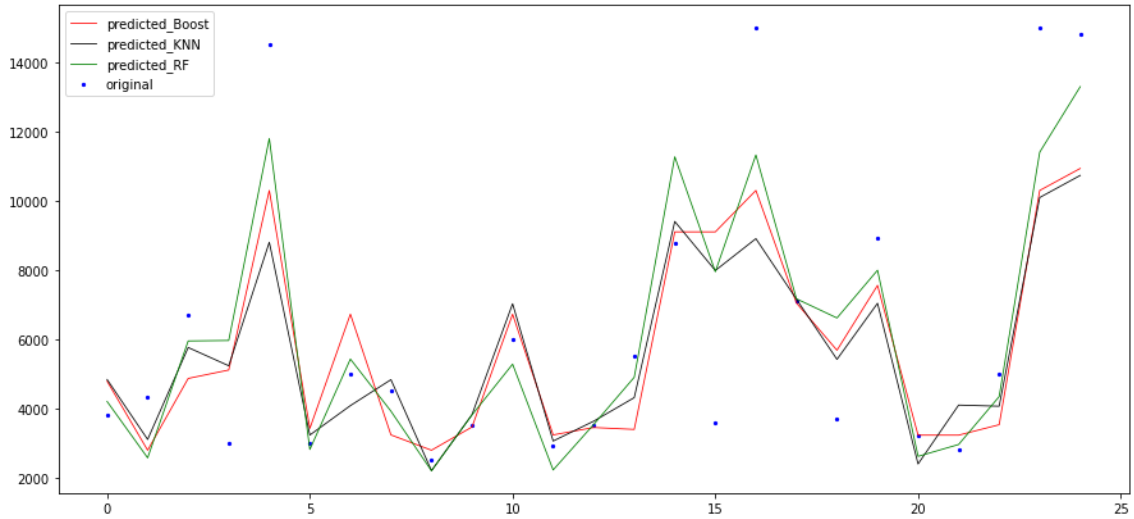
Berikut merupakan hasil dari MSE yang dilakukan oleh ketiga model Machine Learning.



Dapat dilihat pada gambar diatas, algoritma Random Forest memiliki nilai error yang paling kecil (berada di angka 52,6 untuk train dan 237 untuk test) dibandingkan kedua algoritma lainnya. Sedangkan untuk algoritma Boosting memiliki nilai error yang paling besar dibandingkan yang lainnya (berada di angka 317,32 untuk train dan 344,75 untuk test).

Selain itu terdapat juga tabel hasil prediksi model-model serta nilai aktual dan visualisasi plottingan yang menampilkan data test dengan data yang diprediksi oleh model Machine Learning.

y_true	prediksi_RF	prediksi_Boosting	prediksi_KNN
3800.0	4196.1	4781.6	4830.8
4340.0	2568.6	2788.4	3100.0
6700.0	5943.3	4861.2	5753.8
3000.0	5957.2	5104.0	5226.9
14500.0	11788.6	10288.0	8792.3
2999.0	2810.3	3419.5	3226.9
5000.0	5420.5	6716.5	4076.5
4500.0	3916.2	3230.8	4825.8
2500.0	2195.0	2788.4	2199.8
3500.0	3809.1	3460.1	3819.2
6000.0	5273.2	6716.5	7017.3
2900.0	2219.0	3230.8	3053.8
3500.0	3532.1	3444.6	3626.9
5500.0	4894.2	3389.5	4306.9
8750.0	11260.3	9092.3	9392.3
3600.0	7941.1	9092.3	7976.9
15000.0	11311.8	10288.0	8896.2
7100.0	7159.8	7039.5	7149.2
3700.0	6606.5	5674.7	5411.5
8900.0	7985.6	7547.4	7030.8



Titik biru pada plot merupakan data test yang benar, dan garis plot berwarna merah diperuntukkan algoritma Boosting, garis hitam untuk algoritma KNN, dan garis hijau untuk algoritma Random Forest. Pada gambar diatas, dapat dilihat bahwa nilai prediksi untuk setiap algoritma tidak ada yang persis sama pada titik, yang ada hanya diantara diatas titik (artinya prediksi harga rumah lebih tinggi dari nilai aktual) atau dibawah titik (artinya prediksi harga rumah lebih rendah dari nilai aktual). Sebagai contoh, kita mengambil data pada titik ke-5, nilai aktualnya yaitu 2999, dan prediksi nilai yang paling mendekati titik yaitu prediksi model Random Forest di angka 2810,3, sedangkan prediksi model KNN di angka 3226,9 dan prediksi model Boosting di 3419,5.

Kesimpulan

Pada proyek predictive analysis ini, dapat ditarik kesimpulan berdasarkan prediksi harga rumah yang ada di Tebet, Jakarta Selatan dengan menggunakan tiga model regresi Machine Learning, yaitu bahwa diantara Random Forest, K-Neighbors Regressor, dan AdaBoost, algoritma Random Forest lebih baik dibandingkan yang lainnya. Hal ini dapat dilihat dari nilai *Mean Squared Error* (MSE) yang dihasilkan lebih kecil dibandingkan yang lainnya.

Referensi

Febrion Rahayuningtyas, E., Novia Rahayu, F., Azhar, Y., & Artikel, I. (2021). Prediksi Harga Rumah Menggunakan General Regression Neural Network. *Jurnal Informatika,8*(1), 59–66. https://ejournal.bsi.ac.id/ejurnal/index.php/ji/article/view/9036