BAB I PENDAHULUAN

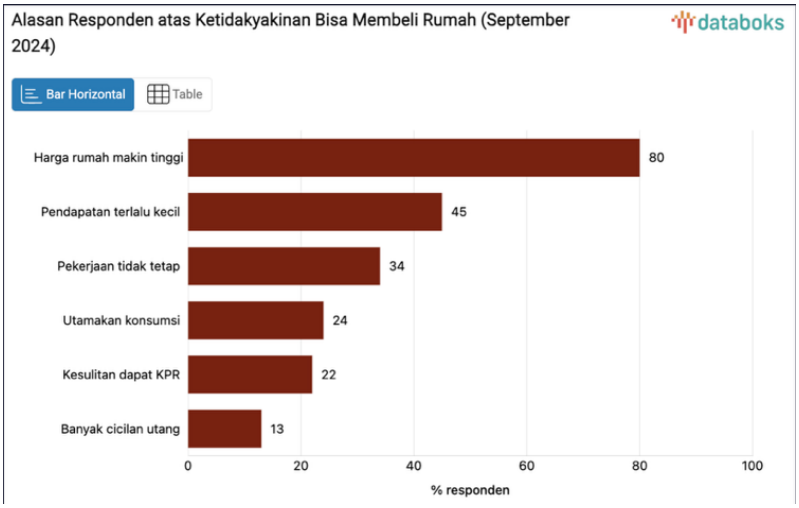
Pada bab pendahuluan ini diawali dengan pemaparan beberapa permasalahan yang menjadi dasar dilaksanakannya penelitian. Uraian tersebut mencakup latar belakang permasalahan yang menjelaskan kondisi serta fenomena yang melatarbelakangi penelitian, batasan masalah yang dibuat untuk memperjelas fokus kajian agar tidak meluas, serta tujuan penelitian yang ingin dicapai. Selain itu, pada bab ini juga akan dijelaskan manfaat penelitian baik secara teoritis maupun praktis, serta ruang lingkup penelitian yang memberikan gambaran mengenai sejauh mana penelitian ini dilakukan. Dengan demikian, bab pendahuluan berfungsi sebagai landasan awal yang komprehensif untuk memahami arah dan kontribusi penelitian secara keseluruhan.

1.1 Latar Belakang

Semakin berkembangnya pengetahuan dan wawasan pada generasi muda, terutama pada Generasi Y atau Generasi Z yang baru memasuki dunia pekerjaan, membawa dampak signifikan pada persepsi atau pandangan mereka terhadap properti, khususnya rumah. Generasi muda ini mulai menyadari pentingnya memiliki hunian sebagai kebutuhan primer sekaligus investasi masa depan. Namun, tantangan ekonomi yang semakin kompleks, seperti kenaikan biaya hidup, fluktuasi pasar, dan keterbatasan penghasilan, membuat mereka harus lebih berhati-hati dalam merencanakan keuangan, termasuk dalam perencanaan membeli rumah (Wadani et al., 2023). Menurut Dewabrata et al. (2023), jumlah unit rumah yang semakin bertambah tahun semakin terbatas sehingga sulit bagi masyarakat untuk mencapai keinginannya dalam membeli sebuah hunian.

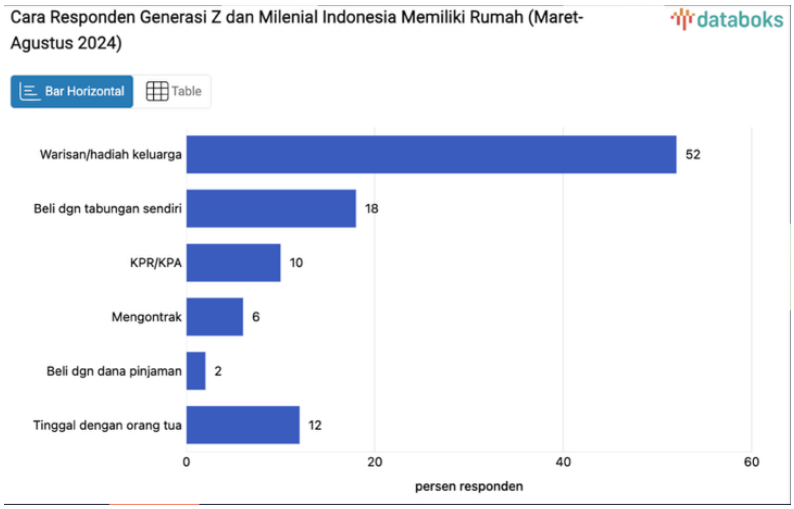
Selain itu, pertumbuhan ekonomi yang kurang stabil turut menyebabkan generasi muda kesulitan mencapai kondisi keuangan yang mapan. Situasi ini berdampak pada kemampuan mereka memenuhi kebutuhan primer dan sekunder, terutama kebutuhan paling penting yaitu memiliki hunian sebagai tempat tinggal jangka panjang.

Ketidakpastian kondisi ekonomi dan fluktuasi harga properti membuat masyarakat membutuhkan informasi yang akurat dan dapat diandalkan terkait harga rumah.



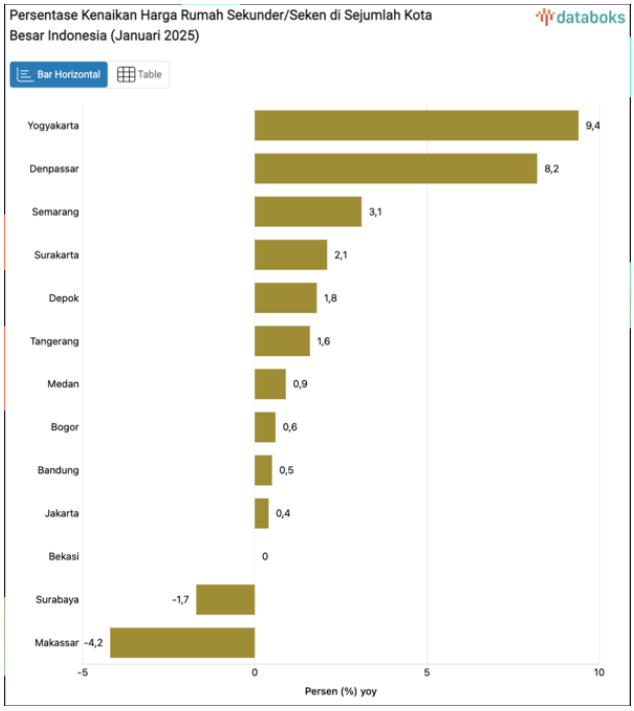
Gambar 1.1 Survey responden membeli rumah

Menurut hasil survey tahun 2024 dari katadata.co.id, berdasarkan gambar 1.1 yang melibatkan 450 responden kelas menengah milenial sebanyak 60% dan gen Z sebanyak 40%. Adapun komposisi gendernya, yakni 53% laki-laki dan 47% perempuan. Kemudian diproleh survey lain dimana bagaimana generasi saat ini memiliki rumah.



Gambar 1.2 Generasi Z dan Milenial memiliki rumah

Kemudian pada gambar 1.2, melibatkan 1.500 responden, terdiri dari 750 responden milenial dan 750 responden gen Z yang tersebar di 12 kota besar Indonesia dengan umur 17-43 tahun. Setelah terdapat survey berdasarkan antar generasi, survey lain menunjukkan kenaikan harga rumah di beberapa jumlah kota besar di Indonesia pada Januari 2025.



Gambar 1.3 Kenaikan harga rumah seken

Berdasarkan gambar 3 yang bersumber dari katadata.co.id, pada januari 2025 menunjukkan jika terjadi beberapa kenaikan harga di awal tahun dengan presentase sebesar 9,4% di Kota Yogyakarta, Denpasar sebesar 8,2% hingga Kota Semarang sebesar 3,1%. Selain itu, pertumbuhan jumlah penduduk yang signifikan, diiringi oleh banyaknya pembangunan rumah oleh pemerintah dan pengembang, menciptakan berbagai pilihan bagi masyarakat untuk memiliki hunian (Fuadah et al. 2020). Namun, dengan semakin banyaknya opsi, konsumen juga menghadapi tantangan dalam menentukan pilihan yang tepat berdasarkan harga rumah yang terus meningkat setiap tahunnya.

Kenaikan harga rumah didorong oleh beberapa faktor, antara lain pertumbuhan jumlah penduduk, meningkatnya permintaan masyarakat akan hunian, ketersediaan lahan yang terbatas, serta kenaikan harga bahan bangunan, dan lain sebagainya. Fenomena ini menciptakan tantangan bagi generasi muda yang baru memulai karir, seperti generasi Z. Mereka perlu mempertimbangkan perekonomian mereka secara matang untuk menghadapi kenaikan harga rumah yang terus berlangsung (Wariyono, 2024).

Maka dari itu, salah satu solusi yang dapat membantu calon pembeli rumah dengan adanya pemanfaatan algoritma machine learning untuk memprediksi harga rumah dengan mempertimbangkan beberapa faktor yang dapat mempengaruhi harga rumah tersebut. Pada penelitian sebelumnya, analisis data terhadap harga rumah dapat memberikan gambaran yang lebih jelas tentang tren harga di masa depan, sehingga mereka dapat membuat keputusan yang lebih selektif dalam pengambilan keptusan. Menurut Fuadah et al. (2020), penggunaan algoritma prediksi tidak hanya membantu menentukan harga rumah yang sesuai dengan kemampuan finansial konsumen, tetapi juga meningkatkan kepercayaan mereka dalam berinvestasi di sektor properti.

Dalam situasi ini, prediksi harga rumah menjadi kebutuhan penting untuk memberikan nilai tambah, baik bagi calon pembeli maupun pelaku investasi. Sebagai bentuk investasi yang menjanjikan, rumah tidak hanya berfungsi sebagai tempat tinggal, tetapi juga aset yang memiliki potensi keuntungan jangka panjang. Oleh karena itu, diperlukan analisis yang teliti untuk memprediksi pergerakan harga rumah dengan akurat agar calon pembeli dapat menghindari potensi kerugian dan memilih properti yang paling menguntungkan (Siregar et al., 2023).

Penelitian ini juga menyoroti pentingnya faktor-faktor yang mempengaruhi harga rumah. Menurut Friedman et al. (2001), keberhasilan suatu model *machine learning* sangat dipengaruhi oleh fitur-fitur input yang digunakan. Fitur-fitur yang dipilih secara tepat dapat menyederhanakan ruang masalah, sehingga membantu meningkatkan daya prediksi algoritma secara signifikan. Berdasarkan dari penelitian sebelumnya faktor terdapat berbagai macam kategori seperti harga rumah, cicilan rumah, lokasi rumah, jenis pemukiman, luas tanah, luas bangunan, kondisi bangunan, jumlah kamar tidur, jumlah

kamar mandi, jumlah garasi dan carport, sertifikat, daya listrik, jumlah lantai bangunan, kondisi properti, dapur, ruang tamu, ruang makan, kondisi perabotan, material bangunan, material lantai, arah rumah, konsep rumah, jangkauan internet, lebar jalan, sumber air, tahun bangunan, tahun renovasi, fasilitas rumah, jarak dari pusat kota. Dengan adanya implementasi algoritma *Random Forest, XGBoost, dan Support Vecto Regression* dengan memprediksi harga rumah dapat dilakukan lebih akurat berdasarkan data yang tersedia (Nuzurialini, 2024). Sehingga penggunaan dataset pada penelitian ini telah disusun sedemikian rupa sesuai dengan unsur pertimbangan dalam membeli sebuah rumah seperti spesifikasi rumah (harga rumah, cicilan rumah, luas tanah, luas bangunan, kondisi bangunan, jumlah kamar tidur, jumlah kamar mandi, jumlah garasi carport, sertifikat, daya listrik, jumlah lantai bangunan, kondisi properti), interior dan exterior (dapur, ruang tamu, ruang makan, kondisi perabotan, material bangunan, arah rumah, konsep bangunan), tentang properti (jangkauan internet, lebar jalan, sumber air, tahun bangunan, tahun renovasi, jarak dari pusat kota) atau pembuatan *feature* baru yang dapat mendukung dataset.

Selain itu, beberapa penelitian sebelumnya kurang objektif dalam menyajikan hasilnya, seperti penggunaan dataset yang kurang representatif terhadap kondisi di lapangan, fokus yang terbatas pada satu metode penerapan, serta evaluasi yang kurang menyeluruh. Oleh karena itu, perlunya suatu penelitian mendalam dengan dukungan dan kerja sama dari pihak-pihak terkait selama pelaksanaan penelitian ini.

Menurut Efendi et al. (2024), Algoritma *Random Forest* menjadi metode yang dapat meningkatkan hasil akurasi dengan kemampuannya untuk menangani overfitting dengan implementasi *bootstrap sampling* yang menggabungkan hasil prediksi rata-rata dari beberapa *decision tree*, sehingga dapat menghasilkan model yang lebih stabil dan akurat. Selain itu, *Random Forest* mampu menangani dataset dengan dimensi tinggi dan variabel yang kompleks, serta dapat mengatasi data yang tidak seimbang. Namun, hasil akurasi yang tinggi dan akurat juga dipengaruhi oleh beberapa hal lain seperti jumlah total dataset serta jenis dan jumlah *feature* yang dimiliki pada dataset penelitian (Widyastuti, 2018). Tidak hanya itu, pengoptimalan seperti pembagian *data training* dan *data testing*, parameter jumlah pohon, *hyperparameter* dan lain-lain saat *modeling* tentu juga

berpengaruh dalam upaya mendapatkan hasil akurasi yang akurat dan efisien. Maka dari itu, dalam upaya mendapatkan hasil akurasi yang akurat perlunya memiliki dataset yang sudah diolah dengan baik serta pengoptimalan *modeling* tepat dan efisien.

Berdasarkan penelitian yang dilakukan oleh Abigail et al. (2021), harga rumah merupakan hal yang penting karena dapat membantu pemilik tanah, pemilik perumahan, dan pembuat kebijakan dalam menghitung valuasi properti dan menentukan harga jual yang wajar. Hal ini dapat membantu pembeli potensial dalam melakukan pembelian rumah pada waktu yang tepat. Sehingga penelitian tersebut melakukan eksplorasi penggunaan teknik *machine learning, Random Forest* untuk memprediksi harga rumah. Hasil dari penelitian tersebut menunjukkan bahwa evaluasi kinerja model menggunakan berbagai metrik seperti *Mean Absolute Error* (MAE), R^2 atau *Coefficient of Determination* menunjukkan bahwa model *Random Forest* memiliki performa yang baik dalam memprediksi harga rumah, dimana model menunjukkan perbedaan prediksi harga rumah dengan margin kesalahan sekitar **±5** dari nilai aktual, yang menunjukkan akurasi yang dapat diterima. Namun sayangnya dalam penelitian tersebut menggunakan data terbatas hanya 506 data *entry* dengan dataset *Boston housing* yang diambil dari *UCI Machine Learning Repository*. Selain itu, karena penelitian ini hanya berfokus pada hasil akhir untuk mencapai akurasi atau kinerja yang baik dalam memprediksi nilai target, seperti harga rumah, sehingga kurangnya pengoptimalan terhadap model seperti *feature engineering, feature importance, model tuning*.

Kemudian menurut Fitri (2023), dengan terus bertambahnya angka penduduk maka semakin meningkatnya kebutuhan tempat tinggal. Dan tidak hanya itu, faktor lainnya juga mempengaruhi harga rumah seperti kondisi fisik mulai dari luas tanah, luas bangunan, kondisi bangunan, serta lokasi. Sehingga dalam penelitiannya memberikan informasi mengenai prediksi harga rumah yang akurat untuk perencanaan masa depan. Penelitian ini melakukan analisis perbandingan hasil prediksi dengan menerapkan 3 Metode. Dimana hasil prediksi tertinggi model *Random Forest Regression* yang memiliki akurasi sebesar 81,5. Tetapi meskipun penelitian ini membandingkan 3 metode yang berbeda, terdapat kurangnya pengoptimalan pada implementasi model tersebut seperti *feature*

engineering, model tuning. Sehingga perlunya penelitian ulang lebih dalam terhadap penggunaan metode tersebut.

Dari berbagai kesulitan yang dialami oleh penjual terhadap pengetahuan serta informasi pada harga rumah mengakibatkan mereka kurang tepat dalam menentukkan harga rumah yang sesuai (Warjiyono et al., 2024). Harga rumah sendiri dipengaruhi oleh banyak faktor, seperti luas tanah, umur properti, jarak ke pusat kota, fasilitas, kualitas properti, lokasi, aksesibilitas, struktur fisik properti, waktu, luas bangunan, jumlah kamar tidur, jumlah kamar mandi, dan garasi (Adetunji et al., 2021). Prediksi harga rumah yang akurat merupakan tantangan besar (Zhan et al., 2023) sehingga perlu dikembangkan model prediksi yang dapat membantu penjual dan pembeli menentukan harga rumah dengan lebih baik.

Menurut Breiman (2001), *Random Forest* merupakan algoritma yang terdiri dari beberapa *decision tree* berdasarkan subset acak dari data dan *feature* untuk meningkatkan akurasi prediksi, sehingga memperoleh keragaman dalam model seperti pengambilan sampel secara acak, kombinasi hasil dengan menggabungkan prediksi dari banyak pohon yang berbeda dan dapat mengurangi risiko *overfitting*. Sering diketahui dalam penelitian harga rumah dataset tentunya memiliki harga yang sangat bervariatif sehingga kemungkinan terdapat *noise* atau *outlier*, dan kemampuan dalam menggeneralisasi yang baik daripada metode lainnya dinilai lebih efektif dalam memprediksi harga rumah yang sering berubah dari waktu ke waktu. Maka dari itu, sesuai dengan tujuan dalam penelitian ini yaitu mendapatkan hasil akurasi yang tinggi, dilakukan pengoptimalan menggunakan metode *Random Forest* dalam memprediksi harga rumah.

Berdasarkan beberapa penelitian sebelumnya, penelitian ini melakukan optimalisasi dengan algoritma *Random Forest* dalam memprediksi harga rumah dengan melakukan beberapa skenario dan implementasi seperti *Data cleaning*, *Exploratory Data Analysis* (EDA) untuk menemukan pola distribusi data, korelasi antara fitur, mendeteksi anomali seperti *outlier*, dan visualisasi yang dapat memahami karakteristik data sebelum digunakan, *feature transformation* seperti *label encoding* dan *scaling* setelah *splitting data*, kemudian melakukan *feature combination* dengan menggabungkan beberapa

variabel independen menjadi *feature* baru untuk meningkatkan akurasi model *regresi*, yang dilanjutkan dengan percobaan pada *feature selection* berdasarkan *feature importance* dimana pada tahap ini membagi 3 skema yaitu menggunakan seluruh *feature*, menggunakan 5 *feature* yang memiliki persentase tertinggi, menggunakan *feature* yang nilai kepentingannya lebih dari 94%, kemudian pada komposisi pembagian *data testing* dan *data training* juga dilakukan 3 skema yaitu dengan ratio 80:20, 70:30, dan 60:40 dimana pembagian data tersebut merupakan standar dalam banyak penelitian karena memberikan hasil yang konsisten dan dapat diandalkan. Selain itu, dengan persentase tersebut dapat menghindari *overfitting* dan menjadi keseimbangan antara *data training* dan *data testing* (Muraina, 2022).

Berdasarkan penelitian sebelumnya, pemilihan ketiga implementasi algoritma dalam penelitian ini karena algoritma tersebut memiliki kelebihan atau keunggulan toleransi yang baik terhadap *noise* dan *outlier*, akurasi prediksi yang tinggi dan tidak rentan terhadap *overfitting* (Zhang et al. 2022). Selain itu, pada algoritma *Random Forest* merupakan *ensemble* yang dapat menghasilkan nilai akurasi lebih tinggi dibandingkan algoritma prediksi lainnya (Hadi et al. 2024). Maka dari itu, penulis memilih menggunakan algoritma *Random Forest Regression* yang sesuai dengan arah tujuan penelitian ini. Menurut Widyaningsih et al. 2021, selain untuk menghindari *overfitting*, *cross-validation* dilakukan untuk memastikan bahwa model tidak hanya bekerja dengan baik di dataset pelatihan, tetapi juga dapat menangani data baru dengan baik.

Menurut Riando et al. (2024), algoritma *XGBoost* unggul karena kemampuannya dalam menangani kumpulan data dengan banyak *feature* dan interaksi yang kompleks antar variabel. Diketahui dalam penelitian ini juga menggunakan dataset dengan jumlah sebanyak 31 *feature* yang saling memiliki korelasi satu sama lain. Implementasi *boosting* bekerja dengan cara melakukan pembelajaran secara berulang, di mana setiap model baru yang dibentuk akan berfokus pada kesalahan (error) yang dilakukan oleh model sebelumnya, sehingga secara bertahap dapat menurunkan tingkat kesalahan secara keseluruhan serta kemampuannya dalam melakukan regularisasi untuk mencegah *overfitting* (Mubarok et al. 2022). Maka dari itu, berdasarkan penelitian sebelumnya menggunakan algoritma *XGBoost* dalam penelitiannya.

Menurut Bastian et al. (2024), Algoritma *Support Vector Regression* (SVR) bekerja dengan cara mencari sebuah garis tengah atau pemisah optimal yang disebut *hyperplane*, yaitu garis yang memiliki jarak atau margin terdekat terhadap pola data yang diamati. Dalam konteks ini, *SVR* tidak hanya berfokus pada pemisahan data, tetapi juga berupaya meminimalkan kesalahan prediksi dengan tetap mempertahankan margin yang disebut sebagai *epsilon* (ε). Margin *epsilon* ini merupakan batas toleransi terhadap kesalahan prediksi, di mana model masih dianggap bekerja dengan baik selama data jatuh di dalam batas ini.

Selanjutnya, dalam proses pembentukan algoritma *SVR*, terdapat beberapa parameter penting yang sangat memengaruhi performa model, seperti parameter C (*penalty parameter*) dimana parameter ini berfungsi untuk mengontrol sejauh mana model memberikan penalti terhadap data yang melampaui margin *epsilon*. Nilai C yang besar cenderung membuat model lebih fokus pada meminimalkan error, tetapi bisa mengurangi kemampuan generalisasi model. Di sisi lain, ketika digunakan fungsi kernel *non-linear* seperti *Radial Basis Function* (RBF), terdapat tambahan parameter yaitu *Gamma*, yang berperan dalam menentukan seberapa jauh pengaruh satu data terhadap data lainnya dalam ruang *feature* yang telah ditransformasi oleh kernel.

Maka dari itu, algoritma *SVR* dapat melakukan prediksi secara akurat meskipun pada data berdimensi tinggi, karena kompleksitas perhitungannya tidak bergantung langsung pada jumlah dimensi input. Dengan kata lain, *SVR* tetap efisien dan memiliki performa komputasi yang baik walaupun diterapkan pada dataset dengan banyak *feature* (Yusuf et al. 2024).

Kemudian *hyperparameter tuning* dalam penelitian ini menggunakan *Random Search* dan *Bayesian Optimization*, dimana percobaan pada keduanya dipilih karena pada *Random Search* dapat melakukan eksplorasi lebih banyak kombinasi dalam waktu singkat, meskipun kombinasi yang dilakukan secara acak, *Random Search* seringkali dapat menemukan kombinasi *hyperparameter* dengan hasil yang optimal (Ramadhan et al. 2024). Dan pemilihan *Bayesian Optimization* karena dapat mencari kombinasi *hyperparameter* yang optimal untuk *Random Forest* secara efisien tanpa harus mencoba

semua kemungkinan kombinasi dengan membangun model probabilistik atau sering disebut *Gaussian Process Regression* yang memprediksi *hyperparameter* mana yang bisa memberikan hasil lebih baik (Tuner et al. 2021). Dimana *Bayesian Optimization* ini memilih kombinasi baru berdasarkan kombinasi sebelumnya untuk memaksimalkan perbaikan model. Misalnya, jika *n_estimators* = 100 dan *max_depth* = 10 memberikan hasil R^2 *Score* = 0.85, algoritma akan mencoba kombinasi di sekitar angka tersebut. Maka dari itu, pada tahap optimasi algoritma penelitian ini menggunakan *Random Search* dan *Bayesian Optimization*. Sehingga penelitian yang dilakukan penulis diharapkan dapat memberikan kontribusi dalam bagaimana memperoleh pemodelan dengan akurasi yang optimal dalam memprediksi harga rumah mengingat dinamika kenaikan yang fluktuatif dari harga rumah, sekaligus membantu generasi muda dalam mengambil keputusan keuangan yang lebih tepat.

1.2 RUMUSAN MASALAH

Berdasarkan uraian dari latar belakang, maka diperoleh rumusan masalah dalam penelitian ini antara lain:

1. Bagaimana penerapan optimasi *feature engineering* untuk meningkatkan akurasi prediksi harga rumah dengan algoritma *Extreme Gradient Boosting*, *Random Forest Regression* dan *Support Vector Regression*?
2. Apakah kombinasi *feature* dan algoritma mampu meningkatkan akurasi secara signifikan dibandingkan *baseline* model tanpa optimasi *feature*?
3. Bagaimana strategi pemilihan langkah yang tepat dalam implementasi penelitian dapat memengaruhi pencapaian akurasi yang optimal dan meminimalkan tingkat kesalahan prediksi?

1.3 TUJUAN PENELITIAN

Berdasarkan uraian dari latar belakang, maka diperoleh tujuan penelitian dalam penelitian ini antara lain:

1. Menganalisa hasil pengaruh optimasi *feature engineering* terhadap performa algoritma *XGBoost*, *Random Forest Regression* dan *Support Vector Regression* dalam Memprediksi Harga Rumah.

2. Memberikan rekomendasi dalam optimasi model algoritma *supervised learning regression* yang paling optimal untuk mendukung pengambilan keputusan dalam prediksi harga rumah

1.4 MANFAAT PENELITIAN

Berikut dengan adanya penelitian ini memperoleh manfaat antara lain:

1. Memberikan wawasan tentang bagaimana pengaruh dilakukannya optimasi *feature engineering* terhadap peforma algoritma *XGBoost, Random Forest Regression* dan *Support Vector Regression*
2. Menjadi sebuah referensi bagi penelitian selanjutnya yang mengimplementasikan upaya optimasi pada *preprocessing* hingga *hyperparameter tuning*.

1.5 RUANG LINGKUP

Berikut merupakan ruang lingkup penelitian ini, antara lain:

1. Dataset yang digunakan merupakan data harga rumah pada 2024-2025 dengan lokasi wilayah Kota Surabaya.
2. Penelitian ini berfokus dalam menganalisa bagaimana pengaruh dari optimasi *feature engineering* terhadap performa Algoritma *XGBoost, Random Forest Regression* dan *Support Vector Regression* dalam prediksi harga rumah.

Halaman ini sengaja dikosongkan