BAB III

METODOLOGI PENELITIAN

3.1 Objek Penelitian

Objek penelitian pada penelitian ini adalah memprediksi harga properti rumah di Kota Tangerang Selatan dengan membagi wilayah di daerah Tangerang Selatan seperti wilayah seperti Ciputat, Pamulang, BSD, Serpong, Alam Sutera, Bintaro, dan Pondok Aren menjadi fokus utama. Data yang digunakan untuk penelitian ini dianalisis dari berbagai platform jual beli rumah. Data penelitian ini diperoleh dari platform website Lamudi bersama dengan OLX Properti.

Data yang dapat diambil dari platform Lamudi adalah harga rumah, lokasi, dan spesifikasi rumah, seperti luas tanah, luas bangunan, jumlah ruangan, jumlah kamar tidur, jumlah kamar mandi, jumlah lantai rumah, dan balkon. Selain mencakup data spesifikasi rumah, data dari platform Lamudi juga dapat mencakup fasilitas lingkungan, seperti pos *security*, taman bermain, dan yang lainnya.

3.2 Metode Penelitian

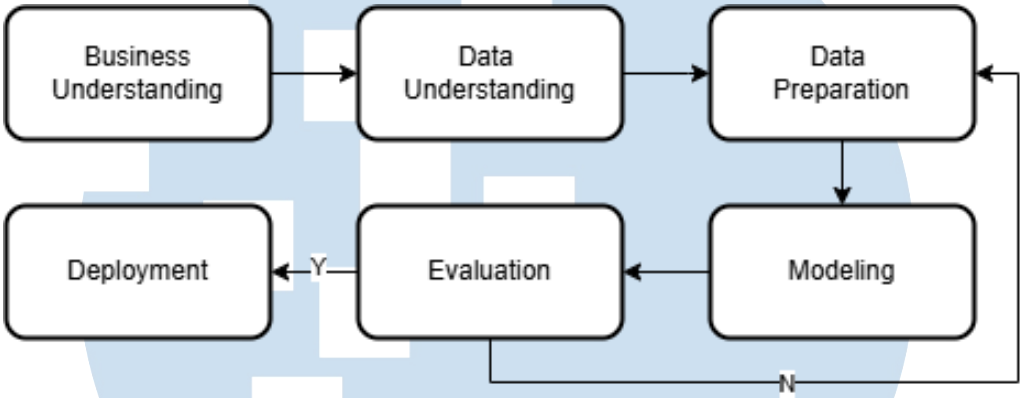
Metode penelitian yang digunakan dalam penelitian ini adalah dengan metode CRISP-DM dengan konsiderasi dari metode lain yaitu KDD untuk melakukan proses *deployment* pada tahap akhir penelitian dan integrasi masalah[26].

Tabel 3. 1 perbandingan Metode *Data Mining*

CRISP-DM	KDD
Memiliki 6 tahap yaitu *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment*.	Memiliki 5 tahap yaitu *selection*, *pre processing*, *transformation*, *data mining*, dan *interpretation*/ *evaluation*.
Memiliki tahap *deployment* di akhir proses.	Tidak memiliki tahap *deployment*, metode KDD diakhiri dengan tahap *evaluation*.
Metode ini dimulai dengan memahami objektif dari penelitian atau projek dengan tahap *business understanding*	Metode ini dimulai dengan membuat *dataset* yang akan digunakan didalam penelitian di tahap *selection*.

Sumber: [26]

Dalam *framework* CRISP-DM terdapat enam fase iteratif yang dimulai pada tahap *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment*. Ilustrasi alur CRISP-DM dapat dilihat di gambar 3.1.



Gambar 3. 1 Proses CRISP-DM

Sumber: [26]

3.2.1 *Business Understanding*

Pada tahap *Business Understanding*, penelitian ini melakukan analisis terhadap industri properti di Indonesia, lalu didalam penelitian ini juga mengumpulan data yang terkait dengan industri properti di Indonesia. Penelitian ini akan mengumpulkan data properti dari kota Tangerang Selatan wilayah Ciputat, Pamulang, BSD, Serpong, Alam Sutera, Bintaro, dan Pondok Aren. Tujuan penelitian ini adalah untuk membuat model prediksi harga rumah di Kota Tangerang Selatan, lalu mengevaluasi performa antara model dengan mengukur performa dari R^2 (*R-Squared*) dan *root mean squared error* (RMSE) dari model *Multiple Linear Regression*, *Gradient Boosting*, dan *Random Forest*. Tahap akhir dari penelitian ini adalah melakukan *deployment* pada model *machine learning* berupa *website*.

3.2.2 *Data Understanding*

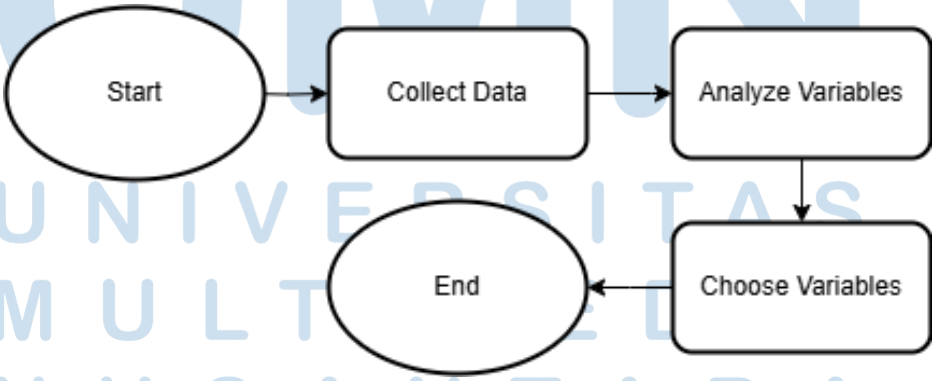
Tahap *data understanding* dilakukan untuk memahami data yang digunakan dalam penelitian. Pada *data understanding,* tahap pertama yang dilakukan adalah mengumpulkan data. Data yang akan digunakan dalam penelitian ini dikumpulkan melalui metode *web scrapping* pada situs jual atau beli rumah yaitu

Lamudi.co.id. Adanya perbandingan antara Website Lamudi.co.id dan Rumah123.com yang dapat dilihat pada Tabel 3.2.

Tabel 3. 2 Perbandingan antara Lamudi dan Rumah123

Kriteria	Lamudi.co.id	Rumah123.com	Keterangan
Kelengkapan Data Properti	Menyediakan detail spesifikasi yang kaya dan terstruktur.	Informasi yang disajikan bisa jadi kurang konsisten antar listing.	Kelengkapan dan konsistensi data di Lamudi.
Struktur Website untuk Web Scraping	Struktur halaman dan elemen HTML yang relatif konsisten.	Struktur website yang sering berubah atau memiliki banyak variasi antar halaman	Kemudahan dalam proses *web scraping* di Lamudi.
Fokus pada Data Spesifik	Menampilkan informasi detail mengenai fasilitas outdoor dan lingkungan properti.	Fokus pada gambar dan deskripsi umum.	Adanya data fasilitas pendukung yang terstruktur di Lamudi.

proses *web scrapping* dilakukan dengan automasi yang bertujuan untuk mengumpulkan *listing* penjualan rumah. Rincian dari tahapan *data understanding* dapat dilihat pada *flow chart* 3.2 berikut.



Gambar 3. 2 *Flowchart* Tahap *Data Understanding*

Selanjutnya, penjelasan dari nama variabel yang akan digunakan dapat dilihat pada tabel 3.3 di bawah.

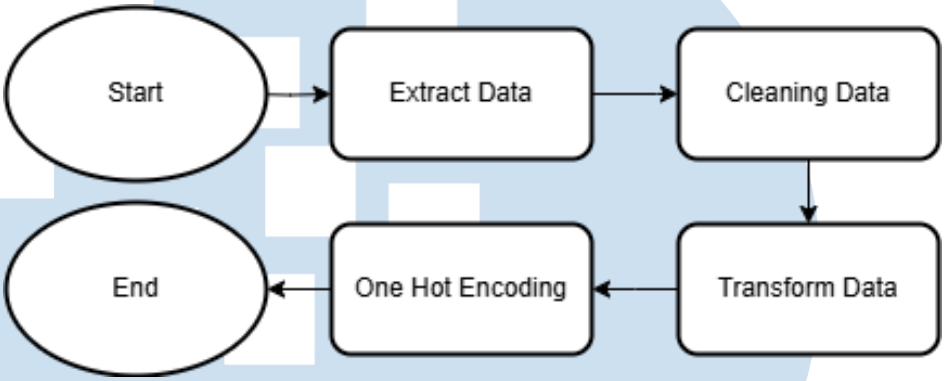
Tabel 3. 3 Variabel Data yang akan digunakan

Fitur	Keterangan
link	*Link* iklan rumah di Lamudi.co.id
title	Judul iklan rumah
location	Lokasi rumah
land_size	Luas tanah rumah
building_size	Luas bangunan rumah
multiple_floor	Jumlah lantai yang dimiliki rumah
bathrooms	Jumlah kamar mandi
bedrooms	Jumlah kamar tidur
carport	Jumlah mobil yang bisa di masukkan ke dalam *carport*
garden	Identifikasi jika rumah memiliki taman
courtyard	Identifikasi jika rumah memiliki halaman belakang
balcony	Identifikasi jika rumah memiliki balkon
security	Identifikasi jika rumah memiliki fasilitas keamanan 24 jam
price	Harga dari iklan rumah

3.2.3 *Data Preperation*

Pada tahap *data preparation*, dilakukannya ekstrasi, seleksi data dan juga mengambil beberapa fitur data yang dibutuhkan didalam penelitian ini. Tahap berikutnya yaitu *cleaning data*, data yang sudah dipilih akan dibersihkan seperti menghapus *outlier*, membuang data-data dan informasi yang tidak diperlukan. Setelah melakukan proses *cleaning data*, selanjutnya melakukan *tranform data* yang bertujuan untuk mengubah tipe data yang tepat seperti untuk kolom *price.* selanjutnya dilakukan proses *One hot encoding* yang bertujuan untuk mengklasifikasikan data lokasi dengan sebuah angka. Dari hasil *data*

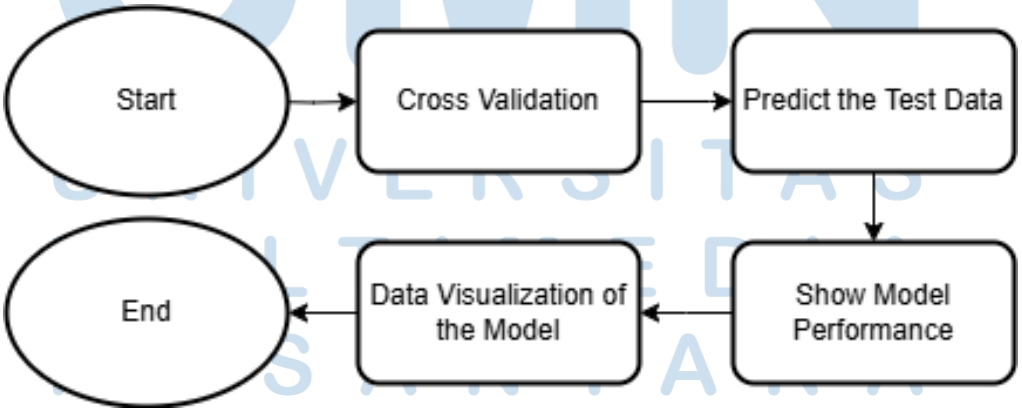
preparation yang dilakukan didapatkan total data sebesar 10307 dengan 18 kolom. Langkah-langkah yang dilakukan dapat dilihat pada gambar *flowchart 3.3* di bawah.



Gambar 3. 3 *Flowchart* Tahap *Data Preparation*

3.2.4 *Modeling*

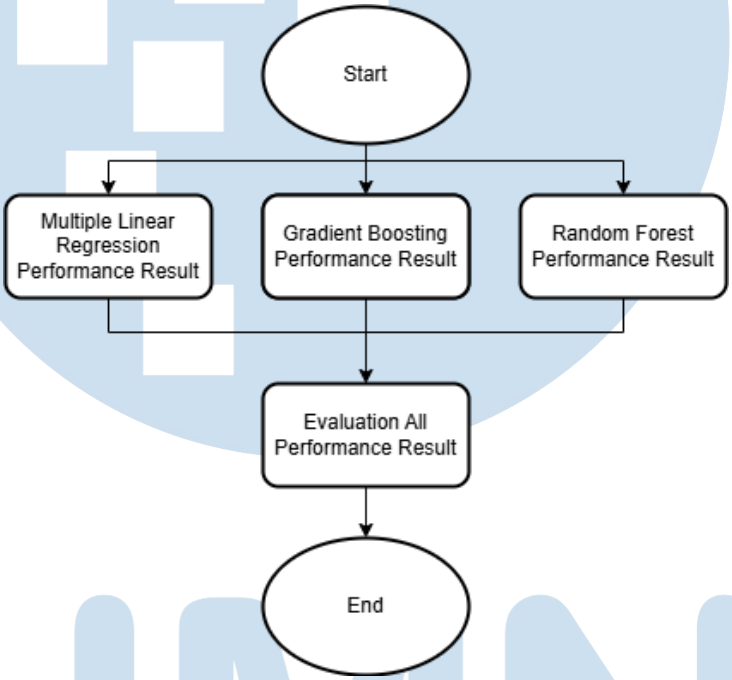
Pada tahap *Modeling,* data yang sudah bersih di tahap sebelumnya akan dilakukan *cross validation* dengan *RandomizedSearchCV*. Penggunaan *RandomizedSearchCV* di penelitian ini dapat meningkatan performa dalam prediksi harga rumah. Penelitian ini menggunakan 5 *fold cross validation*. Setelah melakukan melakukan pemodelan terhadap *Multiple Linear Regression, Gradient Boosting,* dan *Random Forest* akan dilakukannya visualisasi terhadap performa dari ketiga algoritma tersebut, lalu melihat hasil performa manakah yang lebih baik dalam menangani *dataset* penelitian ini. Rincian alur dari tahap *modeling* dapat dilihat pada gambar *flowchart* 3.4 di bawah.



Gambar 3. 4 *Flowchart* Tahap *Modeling*

3.2.5 *Evaluation*

Pada tahap *evaluation*, model algoritma *Linear Regression, Gradient Boosting,* dan *Random Forest* akan di evaluasi untuk melihat hasil performa yang dihasilkan dari ketiga model tersebut. *R-Squared* dan *Root Mean Square Error* (RMSE) merupakan parameter evaluasi yang dilakukan dalam penelitian ini. Hasil dari performa algoritma *Linear Regression, Gradient Boosting,* dan *Random Forest* akan digambarkan dalam bentuk tabel. Rincian alur dari tahap *evaluation* dapat dilihat pada gambar *flowchart* 3.5 di bawah.



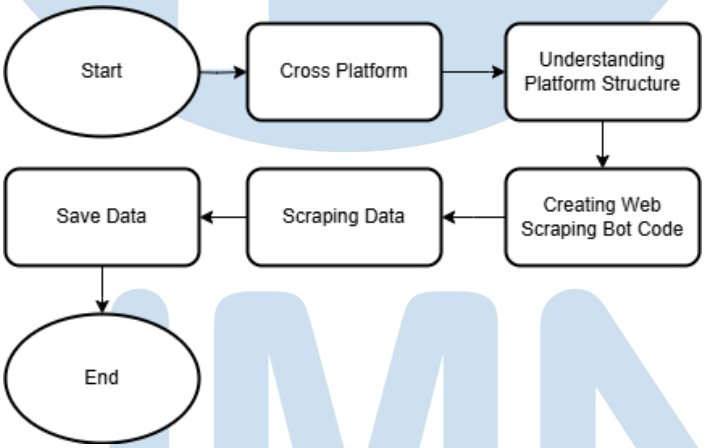
Gambar 3. 5 *flowchart* Tahap *Evaluation*

3.2.6 *Deployment*

Pada tahap terakhir yaitu *deployment*, penggabungan antara rancangan *website* dengan *framework flask* dan model algoritma yang digunakan dalam penelitian ini menghasilkan *tools* yang dapat memprediksi paramater. *Website* yang sudah dibuat dapat digunakan oleh *user* untuk memprediksi harga rumah sesuai dengan kebutuhan dan spesifikasi yang diinginkan.

3.3 Teknik Pengumpulan Data

Tahapan yang dilakukan untuk pengumpulan data dimulai dengan survei situs jual beli properti yang ada di Indonesia sebagai sumber data penelitian. Setelah melakukan survei dan memilih situs jual beli properti, langkah selanjutnya adalah memahami struktur HTML website yang dipilih (Lamudi.co.id) untuk memperoleh informasi-informasi yang akan digunakan, seperti *link*, informasi luas tanah dan luas bangunan rumah. Teknik pengumpulan data menggunakan package bernama *selenium*. *Selenium* akan menarik semua data pada setiap halaman pencarian lalu setelah mendapatkan semua *link* halaman *detail*, dilakukan *scrape* pada semua data disetiap halamannya. Data yang sudah ditarik lalu disimpan ke dalam bentuk *csv* untuk nantinya diproses pada proses selanjutnya. Rincian alur dari teknik pengumpulan data dapat dilihat pada gambar *flowchart* 3.6 di bawah.



Gambar 3. 6 *Flowchart* Teknik Pengumpulan Data

Hasil *web scraping* pada situs Lamudi.co.id menghasilkan data mentah sebanyak 17454 data dan 14 kolom yang berisi data harga rumah yang berlokasi di Kota Tangerang Selatan wilayah Ciputat, Pamulang, BSD, Serpong, Alam Sutera, Bintaro, dan Pondok Aren.

3.4 Teknik Analisis Data

Pemilihan algoritma *Multiple Linear Regression*, *Gradient Boosting*, dan *Random Forest* dalam penelitian ini didasarkan pada strategi untuk membandingkan model dari berbagai tingkat kompleksitas dan karakteristik yang berbeda. *Multiple Linear Regression* dipilih sebagai *base model* yang sederhana untuk menjadi tolak ukur performa dasar dalam melakukan komparasi. Di sisi lain, *Gradient Boosting* dan *Random Forest* merupakan algoritma *ensemble* yang canggih dan dipilih karena kemampuannya yang cenderung memiliki performa lebih baik serta lebih tahan terhadap *overfitting* dibandingkan model linear. Selain itu, ketiga algoritma ini telah terbukti banyak digunakan dan mampu memberikan kinerja yang baik pada penelitian terdahulu dalam kasus prediksi harga rumah, sehingga memungkinkan dilakukannya evaluasi komprehensif untuk menentukan model dengan performa terbaik untuk dataset properti di Tangerang Selatan. Pada tabel 3.4 menunjukkan penjelasan tentang kelebihan dan kekurangan pada setiap algoritma sesuai dengan kriteria.

Tabel 3. 4 Perbandingan Kelebihan dan Kekurangan Algoritma

Kriteria	**Multiple Linear Regression**	**Gradient Boosting**	**Random Forest**	**Support Vector Regression**	**XGBoost**	**K-Nearest Neighbors**
Akurasi Prediksi	Rendah hingga Sedang	Tinggi	Tinggi	Sedang hingga Tinggi	Sangat Tinggi	Rendah hingga Sedang
Kecepatan *Training*	Sangat Cepat	Lambat	Sedang hingga Lambat	Lambat, terutama pada data besar	Cepat (Teroptimisasi)	Tidak ada fase *training* (prediksi lambat)
Penanganan Overfitting	Rentan, tidak memiliki mekanisme internal	Rentan, namun bisa diatasi dengan *tuning* yang cermat	Risiko rendah karena mekanisme *bagging*	Dapat dikontrol dengan baik melalui parameter	Risiko lebih rendah karena memiliki regularisasi internal	Tergantung nilai k, k kecil bisa *overfit*
Kemampuan Menangani Data Non-Linear	efektif untuk hubungan linear	Sangat Baik	Sangat Baik	Sangat Baik, dengan menggunakan *kernel trick*	Sangat Baik	Baik
Kebutuhan *Tuning* Parameter	Hampir tidak ada	Sangat dibutuhkan untuk performa optimal	Dibutuhkan	Sangat dibutuhkan (pemilihan *kernel* dan parameternya)	Sangat dibutuhkan untuk performa maksimal	Kritis
Sensitivitas terhadap *Outlier*	Sangat Sensitif	Cukup Sensitif	Kuat dan tidak terlalu terpengaruh (*robust)*	Cukup Sensitif	Cukup Sensitif	Sangat Sensitif
Interpretasi Model	Sangat Mudah	Sulit	Sedang	Sulit	Sulit	Mudah

Sumber: [40][12][33]