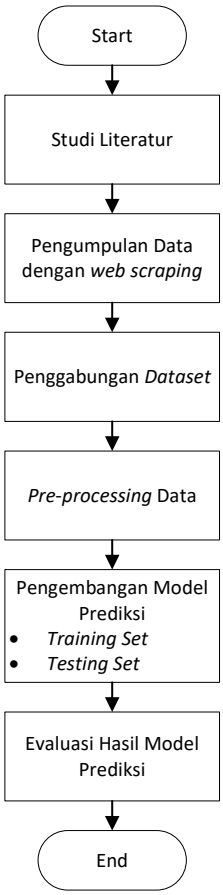
3. METODOLOGI PENELITIAN

Bab ini menjelaskan metodologi yang digunakan dalam penelitian ini, mencakup kerangka penelitian hingga langkah-langkah yang dilakukan untuk mengumpulkan, memproses, dan menganalisis *dataset* menggunakan algoritma ML.

3.1. Kerangka Penelitian

Penelitian ini dilakukan dengan beberapa urutan tahapan yang membentuk sebuah diagram alir seperti yang terlihat pada Gambar 3.1.



Gambar 3.1. Bagan Alir Kerangka Penelitian

Urutan tahapan penelitian ini diawali dengan studi literatur, yang bertujuan untuk mencari ide penelitian dan merangkum permasalahan yang dikutip dari jurnal, buku, dan penelitian-penelitian sebelumnya. Studi literatur ini dilakukan sebagai tahapan awal evaluasi mengenai metode-metode pada penelitian sebelumnya. Tahapan berikutnya adalah pengumpulan data sekunder rumah melalui situs platform *online*.

Setelah data proyek dikumpulkan, dilakukan tahapan pengolahan berupa *data pre-processing* dan akan ditampilkan statistik deskriptif dari *dataset* tersebut. Kemudian dilakukan proses pembentukan model prediksi berdasarkan *dataset training* yang sesuai dengan kebutuhan penelitian untuk menjawab rumusan masalah. Model prediksi tersebut kemudian akan diuji menggunakan *testing set* untuk menguji apakah model prediksi algoritma ML tersebut akurat. Dari hasil pengolahan dan pengujian data tersebut kemudian dapat diambil kesimpulan dari hasil evaluasi model prediksi yang dapat menjawab rumusan masalah dalam penelitian ini.

3.2. Jenis Penelitian

Jenis penelitian yang digunakan adalah penelitian asosiatif kuantitatif. Menurut Sugiyono (2013), penelitian asosiatif adalah penelitian yang bersifat menanyakan hubungan antara dua variabel atau lebih. Selanjutnya menurut Sugiyono, penelitian kuantitatif adalah metode penelitian dengan eksperimen yang konkret, obyektif, terukur, rasional, dan sistematis, karena data penelitian berupa angka-angka dan analisis menggunakan statistik. Jenis penelitian kuantitatif dalam penelitian ini digunakan untuk menjawab tujuan penelitian ini untuk menganalisis tingkat akurasi algoritma pemodelan kecerdasan buatan dalam menilai properti rumah tinggal di Surabaya.

3.3. Data dan Variabel Penelitian

3.3.1. Populasi

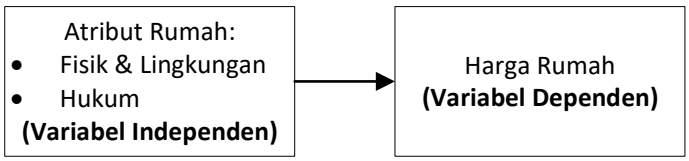
Populasi adalah wilayah generalisasi yang terdiri atas obyek atau subyek yang mempunyai kualitas dan karakteristik tertentu yang ditetapkan untuk dipelajari dan kemudian ditarik kesimpulannya (Sugiyono, 2013). Populasi pada data penelitian ini adalah rumah tinggal di Surabaya

3.3.2. Sampel

Sampel data merupakan bagian dari populasi yang dianggap dapat mewakili populasi yang diteliti. Metode pengambilan sampel menggunakan *non-probability sampling* dengan teknik *purposive sampling*. Menurut Sugiyono (2013), *non-probability sampling* merupakan teknik pengambilan sampel yang tidak memberi peluang yang sama bagi setiap anggota populasi untuk dipilih menjadi sampel. *Purposive sampling* adalah teknik penentuan sampel dengan pertimbangan tertentu. Pada penelitian ini, sampel data yang digunakan adalah rumah tinggal yang ter-*listing* di 31 kecamatan pada *website* Rumah123.

3.3.3. Variabel Penelitian

Pada dasarnya variabel penelitian adalah segala sesuatu yang berbentuk apa saja yang ditetapkan oleh peneliti untuk dipelajari sehingga diperoleh informasi tentang hal tersebut, kemudian ditarik kesimpulannya (Sugiyono, 2013). Variabel bebas pada penelitian ini adalah atribut rumah yang berkaitan dengan fisik dan lingkungan, serta hukum. Sedangkan, untuk variabel terikat pada penelitian ini adalah harga rumah (lihat Gambar 3.2). Faktor ekonomi dan faktor sosial pada penelitian ini tidak diperhitungkan karena data yang digunakan seluruhnya berlokasi di Kota Surabaya, sehingga bersifat homogen dan tidak memiliki pengaruh dalam prediksi harga rumah.



Gambar 3.2. Hubungan Variabel Independen-Dependen

3.3.4. Pengumpulan Data

Pengumpulan data diawali dengan mencari situs platform *online* real estat yang ada tersebar secara publik di internet. Umumnya situs platform seperti itu dimiliki oleh agen properti seperti: Xavier Marks, PropNex Indonesia, Brighton, Ray White Indonesia, Galaxy Property, dll. Selain platform yang berasal dari agen properti, ada beberapa situs properti yang juga menyediakan layanan untuk menampung informasi properti-properti dari beberapa agen, seperti: Rumah123, OLX, 99.co, Lamudi, Rumah.com, dsb. Data rumah yang akan diambil wajib mencakup informasi dasar seperti harga, jumlah kamar tidur, jumlah kamar mandi, luas tanah, luas bangunan, dan lokasi.

Untuk mengetahui situs mana yang paling tepat untuk dilakukan penelitian ini, perlu dilakukan riset pada tiap situs untuk mencari tahu situs mana yang menyediakan informasi dasar dan informasi lainnya yang lebih banyak, serta kemudahan melakukan *scraping* informasi pada situs tersebut. Tabel 3.1 menjelaskan mengenai informasi yang tersedia di 10 situs real estat dan dapat melihat perbandingan kelengkapan informasi tiap situs. Tabel 3.2 menjelaskan mengenai deskripsi dari kolom informasi pada Tabel 3.1.

Tabel 3.1.
Perbandingan Situs *Online* Properti

No	Informasi	Situs								
		XV	PN	BR	RW	GP	R12	OLX	99	LM
1	Lokasi (Kecamatan)	✓	✓	✓	✓	✓	✓	✓	✓	✓
2	Luas Tanah	✓	✓	✓	✓	✓	✓	✓	✓	✓
3	Luas Bangunan	✓	✓	✓	✓	✓	✓	✓	✓	✓
4	Dimensi Bangunan	✓	✓							
5	Kamar Tidur	✓	✓		✓	✓	✓	✓	✓	✓
6	Kamar Mandi	✓	✓		✓	✓	✓	✓	✓	✓
7	Hadap	✓	✓	✓		✓	✓		✓	
8	Sumber Listrik	✓					✓			
9	Jumlah Lantai	✓	✓				✓	✓	✓	
10	Ruang Makan						✓			
11	Ruang Tamu						✓			
12	Kondisi Perabotan		✓	✓		✓	✓		✓	
13	Material Bangunan			✓			✓			
14	Material Lantai			✓			✓			
15	Sertifikat		✓	✓	✓		✓		✓	
16	Garasi						✓			
17	*Carport*				✓				✓	
18	Konsep dan Gaya Rumah						✓			
19	Pemandangan						✓			
20	Terjangkau Internet						✓			
21	Lebar Jalan						✓			
22	Tahun Dibangun						✓		✓	
23	Tahun di Renovasi						✓			
24	Sumber Air					✓	✓			
25	*Hook*						✓			
26	Kondisi Properti						✓			
27	Telepon	✓								
28	Deskripsi	✓	✓	✓	✓	✓	✓	✓	✓	✓
29	Terakhir diperbaharui		✓				✓			

Keterangan pada kolom situs Tabel 3.1 adalah sebagai berikut:

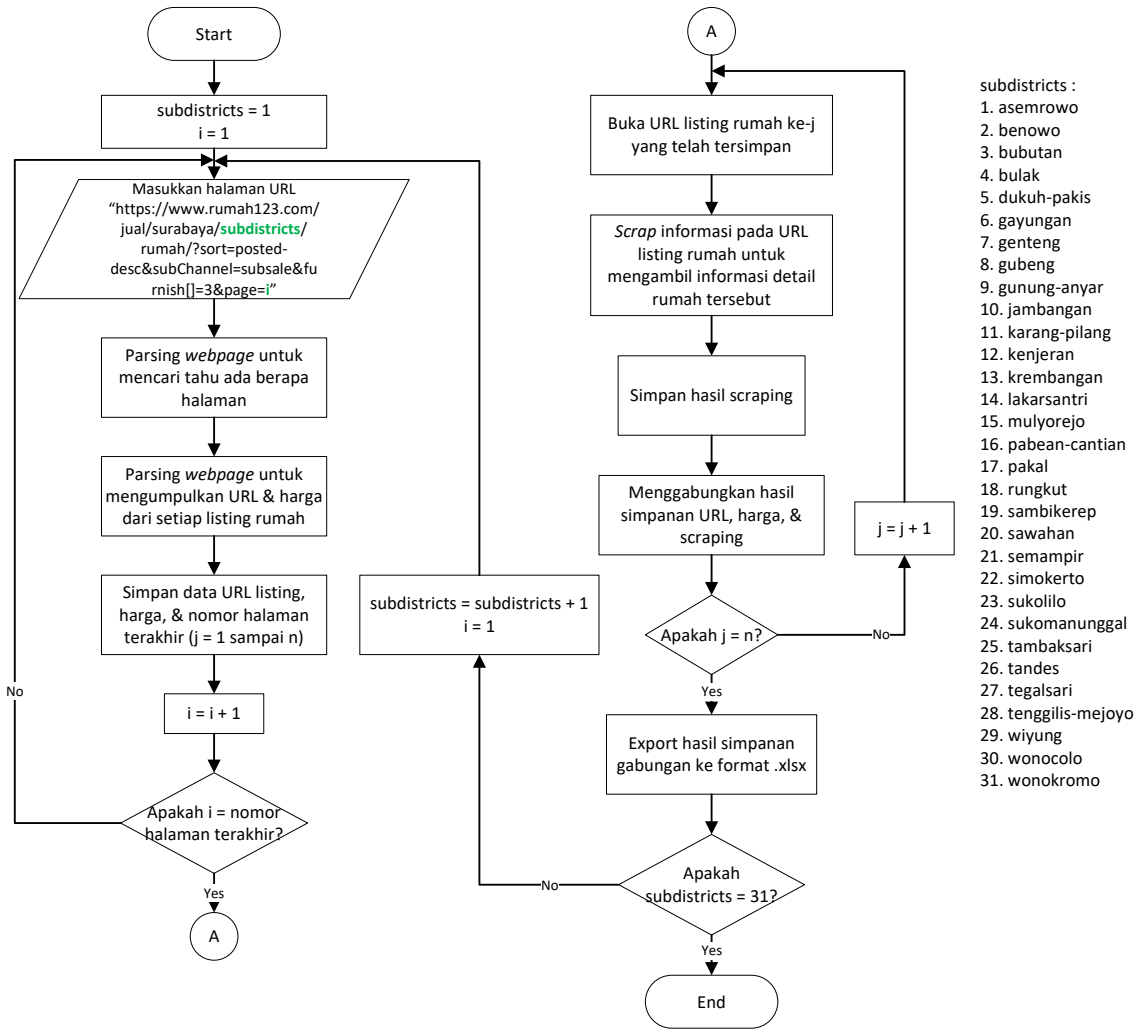
- XV : Xavier Marks
- PN : PropeNex Indonesia
- BR : Brighton
- RW : RayWhite Indonesia
- GP : Galaxy Property
- R12 : Rumah123
- OLX : OLX
- 99 : 99.co
- LM : Lamudi
- R.C : Rumah.com

Tabel 3.2.
Deskripsi Informasi

No	Informasi	Keterangan
1	Lokasi	Lokasi tidak langsung dari *listing* rumah (kecamatan, kode pos, dll)
2	Luas Tanah	Luasan tanah dalam satuan meter persegi
3	Luas Bangunan	Luasan total bangunan dalam satuan meter persegi
4	Dimensi Bangunan	Panjang x lebar rumah dalam satuan meter
5	Kamar Tidur	Jumlah kamar tidur
6	Kamar Mandi	Jumlah kamar mandi
7	Hadap	Arah hadap rumah (utara, selatan, timur, barat)
8	Sumber Listrik	Daya listrik PLN rumah
9	Jumlah Lantai	Total lantai rumah
10	Ruang Makan	Apakah memiliki ruang makan atau tidak
11	Ruang Tamu	Apakah memiliki ruang tamu atau tidak
12	Kondisi Perabotan	*furnished*, *semi-furnished*, atau *unfurnished*
13	Material Bangunan	Batu bata merah, bata ringan, dll.
14	Material Lantai	Granit, Keramik, dll.
15	Sertifikat	SHM, SHGB, dll.
16	Garasi	Jumlah garasi di rumah
17	*Carport*	Jumlah mobil yang dapat dimuat pada halaman rumah
18	Konsep dan Gaya Rumah	Minimalis modern, klasik, dll.
19	Pemandangan	Perkotaan, taman, fasum, dll.
20	Terjangkau Internet	Ya atau tidak
21	Lebar Jalan	Lebar jalan depan rumah
22	Tahun Dibangun	Tahun rumah dibangun
23	Tahun di Renovasi	Tahun rumah direnovasi
24	Sumber Air	PDAM atau bor sumur
25	*Hook*	Apakah rumah berada di posisi *hook* atau tidak
26	Kondisi Properti	Bagus, sudah direnovasi, dll.
27	Telepon	Apakah tersambung dengan telepon rumah atau tidak
28	Deskripsi	Apakah punya kolom khusus untuk *lister* dapat menuliskan deskripsi rumah atau tidak
29	Terakhir diperbaharui	Tanggal *listing* rumah diperbaharui

Setelah melakukan riset, pengumpulan data diputuskan dengan melakukan *web scraping* pada *website* Rumah123 (https://www.rumah123.com). Rumah123 adalah *platform online* real estat yang menyediakan informasi tentang berbagai tipe properti rumah, baik itu rumah baru, bekas, maupun sewa. Rumah123 menyediakan banyak informasi mengenai suatu properti. Tidak hanya informasi wajib dasar, namun ada beberapa informasi tambahan lain juga yang disediakan seperti pada Tabel 3.1.

Informasi lain yang tersedia pada *website* juga akan diambil yang berkaitan dengan atribut dan karakteristik dari properti tersebut seperti umur bangunan, deskripsi *listing*, sertifikat (SHM, SHGB, dll.), dsb. Untuk melakukan *scraping* pada situs Rumah123, diperlukan perencanaan proses *web scraping* yang akan membantu dalam proses pengambilan informasi secara otomatis. Gambar 3.3 menjelaskan tentang perencanaan alur proses *web scraping*.



Gambar 3.3. Bagan Alir Perencanaan Pengumpulan Data

Pertama-tama untuk dapat mengotomatiskan proses pengambilan data, perlu mengetahui pola dari URL situs Rumah123 dengan mencoba memasukkan lokasi kota Surabaya pada kotak pencarian. Diketahui bahwa pola dari situs Rumah123 untuk mencari *listing* rumah tinggal yang dijual di kota Surabaya adalah menggunakan URL seperti pada bagan alur. Apabila "subdistricts" diisi dengan "asemrowo" dan "i" diganti dengan "1", maka URL tersebut akan menunjukkan dan memfilter *listing* rumah tinggal yang berada di kota Surabaya, kecamatan Asemrowo pada halaman 1. Untuk berpindah ke halaman berikutnya, cukup dengan mengganti "i" dengan angka 2 dan seterusnya sampai ke halaman terakhir *listing*.

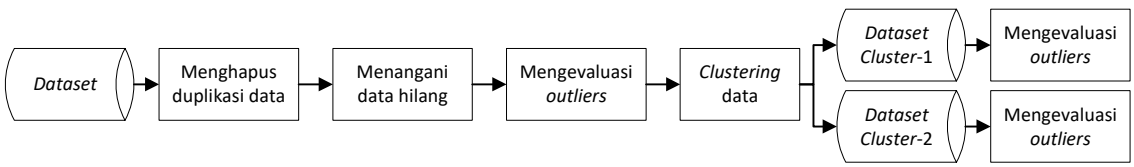
Setelah mengetahui pola tersebut, kemudian dapat dilanjutkan dengan mengambil informasi mengenai *link URL* dan harga tiap rumah di tiap halaman, dimulai dari halaman 1 hingga halaman terakhir pada *listing* rumah di kecamatan Asemrowo. Kemudian, informasi URL dan harga akan disimpan dan URL yang telah tersimpan akan dibuka satu per satu secara otomatis dengan metode *web scraping* yang kemudian akan disimpan dan diekspor menjadi *file* berformat *spreadsheet*. *File spreadsheet* inilah yang akan menjadi *dataset* pada penelitian ini. Setelah kecamatan Asemrowo telah selesai, langkah yang serupa akan diulangi untuk kecamatan Benowo sampai kecamatan wonokromo.

3.4. Penggabungan *Dataset*

File spreadsheet yang berisi informasi dari tiap *listing* rumah di masing-masing kecamatan Surabaya akan digabungkan menjadi satu *file spreadsheet*. Hal ini bertujuan supaya proses berikutnya dapat dilakukan *pre-processing* data.

3.5. *Pre-processing* Data

Tahap ini bertujuan untuk memastikan bahwa data yang digunakan dalam pelatihan dan pengujian model ML berada dalam kondisi optimal, sehingga dapat meningkatkan akurasi dan kinerja model. Proses *pre-processing* data meliputi beberapa langkah seperti pada Gambar 3.4, yaitu: 1) menghapus duplikasi data, 2) menangani data hilang, 3) *clustering* data, dan 4) mengevaluasi *outliers*.



Gambar 3.4. Bagan Alir Perencanaan *Pre-processing* Data

3.5.1. Menghapus Duplikasi Data

Penghapusan duplikasi data dilakukan dengan melakukan filter pada *dataset* yang memiliki nilai sama persis antara data satu dengan data lainnya, sehingga tidak ada lagi data yang memiliki nilai sama persis di setiap variabelnya. Penghapusan duplikasi ini bertujuan untuk menghindari bias ketika dilakukan proses *training*.

3.5.2. Menangani Data Hilang (*Missing Values*)

Di setiap data yang diambil ketika melakukan *web scraping*, tidak semuanya memiliki nilai pada variabelnya. Untuk variabel independen yang tingkat kepentingannya tidak signifikan, maka variabel pengamatan yang memiliki persentase data hilang di atas lima puluh persen (>50%) akan dihapus. Hal serupa juga telah dilakukan oleh Quang (2020) pada *dataset* penelitiannya yang berisi lebih dari 300.000 pada langkah untuk menginvestigasi data hilang, di mana Quang menghapus variabel yang memiliki data hilang dengan persentase di atas 50% (Quang, Nguyen, Dang, & Mei, 2020). Selanjutnya, dalam satu baris yang masih memiliki data hilang dilakukan penanganan dengan cara menghapus keseluruhan data dalam satu baris tersebut.

3.5.3. Clustering

Proses *clustering* ditujukan untuk menemukan dan mengelompokkan data berdasarkan pola-pola tertentu atau karakteristik yang dimiliki masing-masing data, untuk mendukung pembelajaran ketika proses *training*. Hal ini dilakukan karena data mentah yang diperoleh akan mengandung banyak tipe rumah karena penelitian tidak hanya dilakukan pada rumah baru, namun juga rumah lama (*second-hand*) yang memiliki beragam informasi dan karakteristik berbeda-beda. Proses *clustering* data menggunakan algoritma *k-means* yang tersedia pada aplikasi IBM SPSS Modeler 18.0.

3.5.4. Mengevaluasi *Outliers*

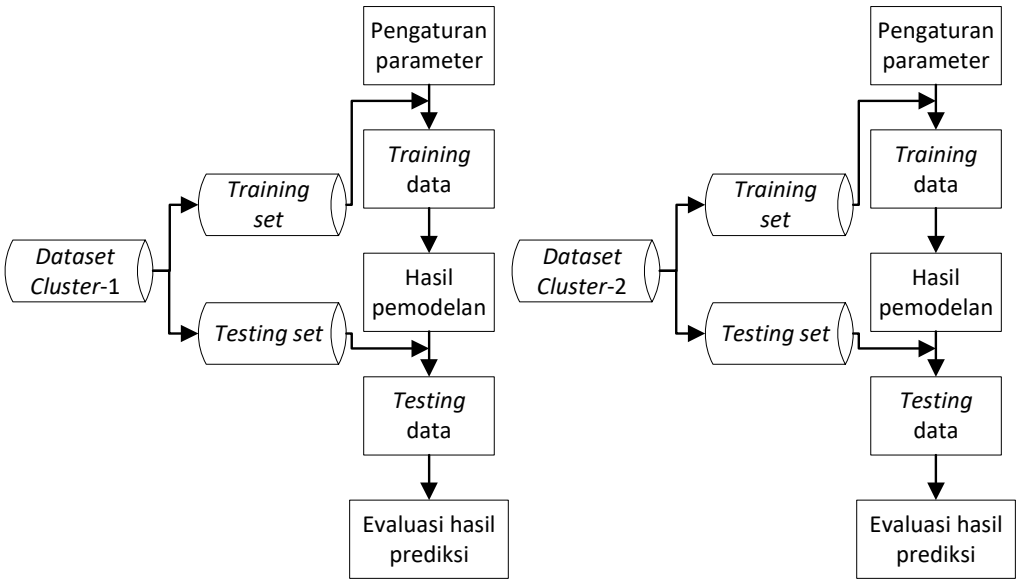
Mengevaluasi *outlier* dilakukan dengan melakukan analisa visual pada *dataset* yang dimiliki dalam bentuk *scatter plot*. Dalam *scatter plot*, *outliers* akan tampak sebagai titik yang jauh dari kumpulan data utama. Aplikasi IBM SPSS Modeler 18.0 memiliki fitur untuk membantu mendeteksi dan memvisualkan *outlier* ketika nilai standar deviasi bernilai 3 dari rata-rata (*mean*) dan nilai *extremes* 5 dari *mean*. Dengan mengevaluasi *outliers* ini maka diharapkan dapat meningkatkan kinerja model statistik dan model prediksi.

3.6. Pengembangan Model Prediksi

Pengembangan model prediksi dilakukan menggunakan aplikasi IBM SPSS Modeler 18.0. Sebelum dilakukannya Gambar 3.5 menunjukkan proses analisa data dan pengembangan model dari penelitian.

3.6.1. *Data Splitting*

Langkah pertama yang dilakukan yaitu memasukkan *dataset* yang telah dipisah menjadi dua *cluster* dalam format *spreadsheet*, untuk kemudian diunggah ke dalam program IBM SPSS Modeler 18.0 (lihat Gambar 3.5). *Dataset* yang telah terunggah kemudian akan dibagi secara acak dengan komposisi 70:30. Gholamy (2018) menyebutkan bahwa sebetulnya semakin banyak data poin maka semakin akurat pula estimasi yang akan dihasilkan. Hal tersebut merupakan ide yang bagus apabila benar-benar yakin bahwa model atau data yang digunakan cukup menggambarkan fenomena yang sesuai (Gholamy, Kreinovich, & Kosheleva, 2018). Namun dalam praktiknya, sering kali data yang dimiliki tidak benar-benar memadai (Gholamy, Kreinovich, & Kosheleva, 2018). Dalam situasi seperti itu, menggunakan semua data yang tersedia untuk menentukan parameter model sering kali akan menghasilkan *overfitting data* (Gholamy, Kreinovich, & Kosheleva, 2018). Analisis empiris telah menunjukkan bahwa hasil terbaik diperoleh jika mengalokasikan 20-30% titik data asli untuk *testing set*, dan menggunakan 70-80% sisanya untuk *training set* (Gholamy, Kreinovich, & Kosheleva, 2018).



Gambar 3.5. Bagan Alir Proses Analisa Data dan Pengembangan Model Prediksi

3.6.2. Pengaturan Parameter

Setelah *training* dan *testing set* terbagi, dilakukan pengaturan parameter untuk menemukan model prediksi yang paling akurat untuk setiap algoritma ML (ANN, SVM, dan CART). Pengaturan parameter dilakukan menggunakan metode *grid search*. *Grid search* adalah metode tradisional yang digunakan untuk *tuning* parameter dalam ML. Metode ini mencoba setiap kombinasi nilai parameter yang disediakan untuk menemukan model terbaik. Metode ini juga digunakan pada penelitian oleh Zhang et al. (2022).

Tabel 3.3.
Rentang Nilai Parameter

Metode	**Parameter**	**Nilai**
ANN	Jumlah *hidden layer*	1∼2
	Jumlah neuron pada setiap *hidden layer*	*automatic*, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100
SVM	*Regularization parameter* (C)	10, 20, 30, 40, 50, 60, 70, 80, 90, 100
	Radial Basis Function (RBF) *gamma*	0.1, 0.5, 1.0
CART	*Minimum records in parent branch* (%)	2, 3, 4, 5, 6, 7, 8
	Minimum records in child branch (%)	1, 2, 3, 4, 5, 6, 7

Nilai pada Tabel 3.3 merupakan nilai parameter yang akan digunakan dan yang akan dibandingkan hasilnya setelah itu pada pengujian untuk melihat parameter dengan nilai mana yang memberikan hasil terbaik. Setelah dilakukan pengaturan parameter, akan dilakukan proses *training* pada 70% data yang akan menghasilkan model prediksi. Hasil model prediksi tersebut akan diuji coba menggunakan *testing set*. Proses *testing* merupakan proses yang mengimplementasikan hasil model prediksi *training* pada 30% data sisanya.

3.7. Pengujian Hasil Model Prediksi

Setelah dilakukan testing, hasilnya kemudian akan dibandingkan menggunakan evaluasi hasil prediksi untuk melihat algoritma dan parameter mana yang dapat menghasilkan hasil yang paling akurat. Penilaian atau pengujian hasil akan menggunakan metode: 1) *linear correlation*, 2) MAE, 3) MAPE, dan 4) RMSE.

3.7.1. *Linear Correlation*

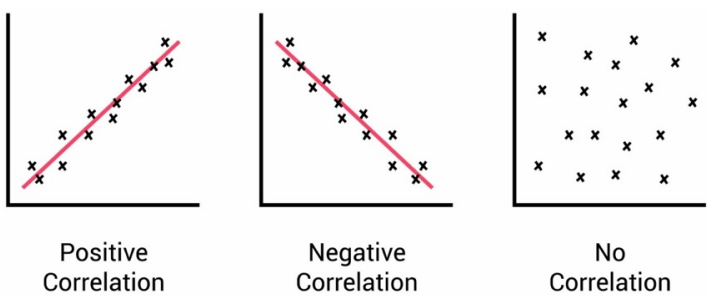
Linear correlation (R) mengukur kekuatan dan arah hubungan linear antara dua variabel. Dalam konteks evaluasi model AI, R mengukur seberapa baik prediksi model berkorelasi dengan nilai aktual. Nilai R dirumuskan dengan formula di bawah ini:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}} \quad \text{.....(3.1)}$$

Di mana:

- r = *linear correlation*
- n = jumlah data
- x = variabel independen
- y = variabel dependen

Nilai R memiliki nilai yang berkisar antara -1 dan 1. Semakin mendekati 1 atau -1, semakin kuat hubungan linear antara prediksi dan nilai aktual. R bernilai 1 berarti hubungan positif sempurna, R bernilai -1 berarti hubungan negatif sempurna, sedangkan R bernilai 0 berarti tidak ada hubungan linear (lihat Gambar 3.6).



Gambar 3.6. Korelasi Linear
Sumber: McLeod, S. (2019). Correlation. Simply Psychology. https://www.simplypsychology.org/correlation.html

3.7.2. *Mean Absolute Error*

Mean Absolute Error (MAE) adalah salah satu evaluasi yang mengukur rata-rata dari semua kesalahan absolut antara nilai prediksi dan nilai aktual. Kesalahan absolut adalah nilai absolut dari selisih antara nilai aktual dan nilai prediksi. MAE memberikan gambaran tentang seberapa besar kesalahan yang dibuat oleh model dalam prediksinya. Formula perhitungna MAE adalah:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \quad \text{.....(3.2)}$$

Di mana:

- n = jumlah data
- x_i = nilai aktual ke-i
- y_i = nilai prediksi ke-i

Semakin rendah nilai MAE, semakin baik kinerja model dalam hal prediksi. MAE yang rendah menunjukkan bahwa kesalahan prediksi rata-ratanya kecil. MAE juga memiliki satuan yang sama dengan variabel dependen yang diprediksi, sehingga mudah diinterpretasikan, serta tidak terlalu sensitif terhadap data *outliers* (nilai ekstrim). *Outliers* adalah data atau observasi yang sangat berbeda dari mayoritas data lainnya dalam suatu *dataset*. Mereka adalah nilai yang berada jauh dari distribusi data utama dan dapat mempengaruhi analisis statistik serta hasil dari model AI.

3.7.3. *Mean Absolute Percentage Error*

Mean Absolute Percentage Error (MAPE) adalah metode evaluasi yang mengukur rata-rata kesalahan absolut sebagai persentase dari nilai aktual. MAPE memberikan gambaran tentang seberapa besar kesalahan prediksi model dalam bentuk persentase, yang memudahkan perbandingan kesalahan prediksi pada berbagai skala data. Nilai MAPE yang rendah menunjukkan model dengan prediksi yang lebih akurat. Rumus untuk menghitung MAPE yaitu:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - y_i}{x_i}\right| \times 100\% \quad \text{.....(3.3)}$$

3.7.4. *Root Mean Squared Error*

Root Mean Squared Error (RMSE) adalah metode evaluasi yang mengukur akar dari rata-rata kesalahan kuadrat antara nilai prediksi dan nilai aktual. RMSE memberikan gambaran seberapa besar kesalahan prediksi model secara rata-rata dalam satuan yang sama dengan variabel dependen. Semakin rendah nilai RMSE, semakin baik kinerja model dalam hal prediksi. RMSE yang rendah menunjukkan bahwa kesalahan prediksi rata-rata kecil. Karena menggunakan kesalahan kuadrat, RMSE lebih sensitif terhadap *outliers* dibandingkan MAE. Rumus untuk menghitung RMSE adalah sebagai berikut:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \quad \text{.....(3.4)}$$