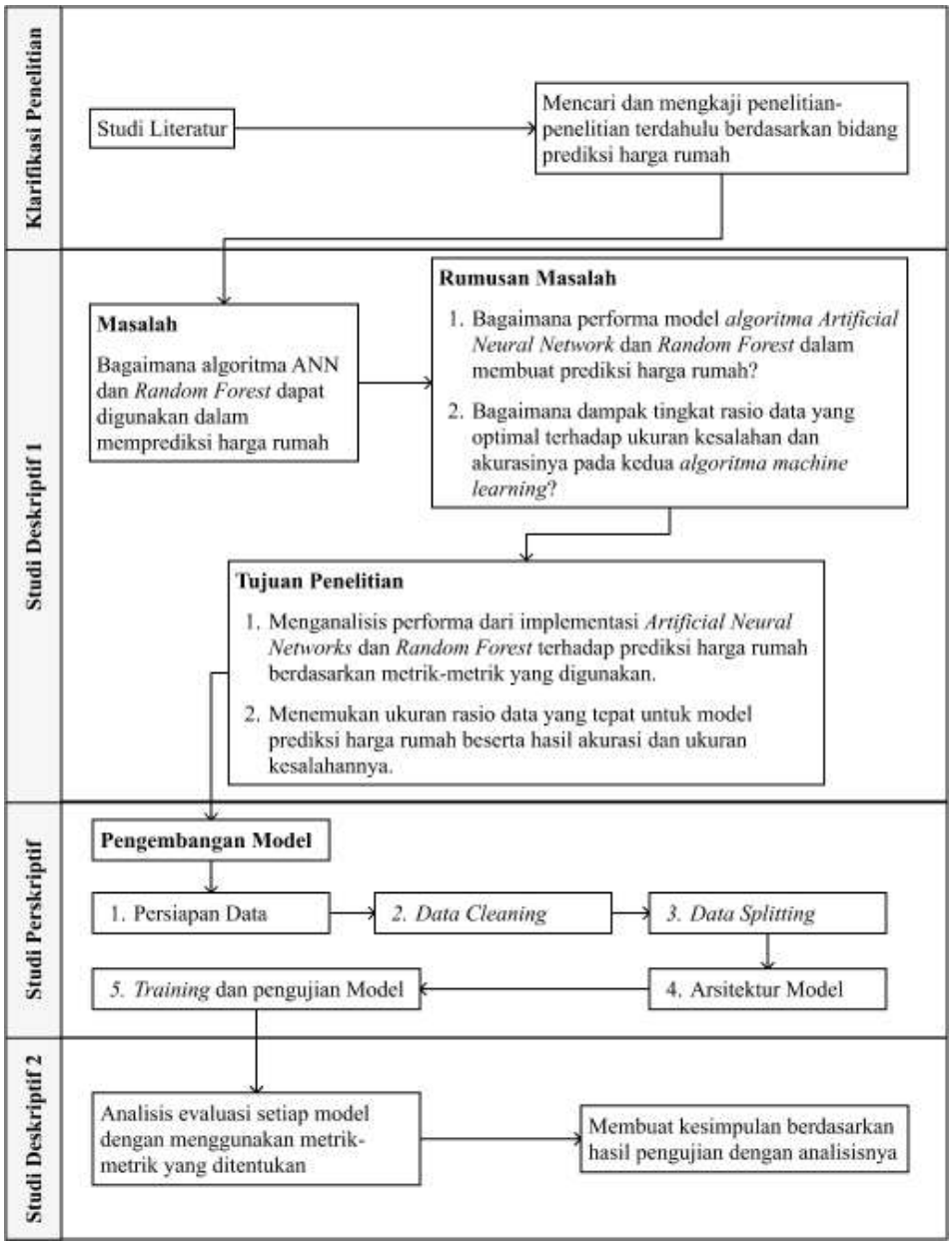
BAB III
METODE PENELITIAN

3.1 Desain Penelitian



Gambar 3.1 Alur DRM

Penelitian ini dilakukan dengan metode penelitian *Design Research Methodology* (DRM). Metode ini memiliki empat tahapan utama dari klarifikasi

penelitian, studi deskriptif 1, studi perskriptif, sampai studi deskriptif 2 (Lattanzio dkk., 2019). Sebagaimana **Error! Reference source not found.** dan **Error! Reference source not found.** menjelaskan mengenai alur penelitian DRM.

3.1.1 Klarifikasi Penelitian

Pada tahap ini, dilakukan kajian literatur dari penelitian sebelum-sebelumnya terkait penelitian *machine learning* dan prediksi harga rumah dengan tujuan untuk mengidentifikasi tujuan dan inti permasalahan dari penelitian. Tidak hanya itu, kajian literatur juga bertujuan agar penelitian yang akan dilakukan masih sejalan dengan penelitian-penelitian sebelumnya dengan mencari teknik yang efektif, rekomendasi dari penelitinya, dan mencari pembaruannya. Selain itu, tahap ini juga dilakukan untuk mencari referensi-referensi yang dapat membantu pada penelitian ini.

3.1.2 Studi Deskriptif 1

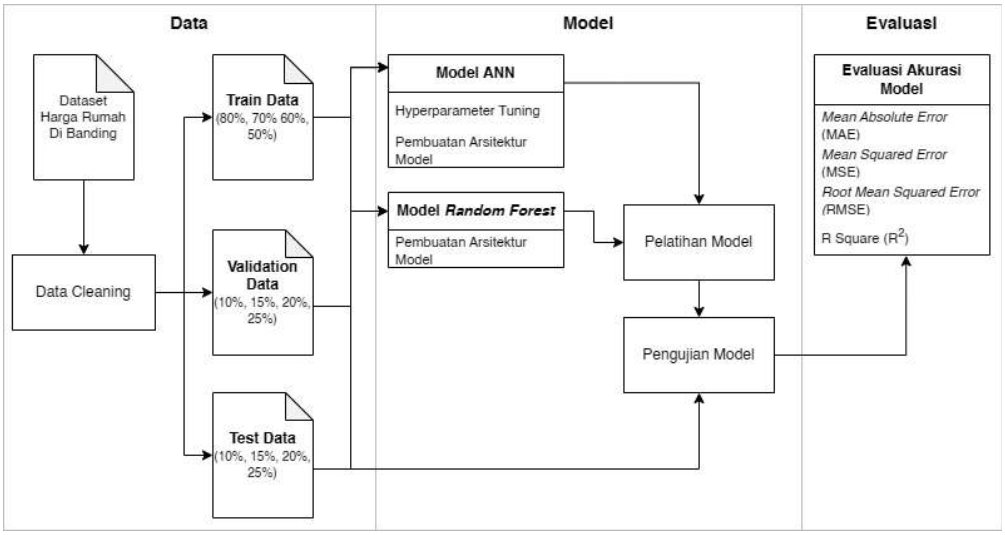
Tahapan ini dilakukan untuk memetakan permasalahan dan tujuan yang didapat setelah melakukan studi literatur pada tahap klarifikasi penelitian sehingga penelitian yang dilakukan tetap sesuai dengan kondisi penelitian yang sudah ada. Permasalahan dan tujuan ditentukan berdasarkan kajian literatur yang sebelumnya dilakukan pada tahap klarifikasi penelitian.

Tahap studi deskriptif 1 dilakukan untuk menentukan tolak ukur dalam mengukur kesuksesan dari penelitian. Tahap ini bertujuan untuk memberikan gambaran mengenai konteks penelitian dan variabel penelitian yang akan digunakan.

3.1.3 Studi Perskriptif

Sebagaimana yang ditampilkan pada Gambar 3.2, ada tiga kegiatan utama dalam studi perskriptif ini: Pengolahan data, pembuatan model, dan evaluasi hasil. Pengolahan data meliputi kegiatan mempersiapkan *dataset* harga rumah, *data cleaning*, dan *data splitting*. Selanjutnya pengembangan model yang meliputi kegiatan pembuatan dan optimasi model, pelatihan model, dan pengujian model yang sudah dibuat. Terakhir adalah kegiatan evaluasi model yang dilakukan dengan menganalisis dan membandingkan hasil-hasil dari setiap model yang sudah dibuat.

Kegiatan evaluasi model juga dilakukan dengan memperhatikan satuan metrik-metrik yang digunakan pada penelitian ini. Metrik-metrik yang digunakan pada penelitian ini dibahas pada bagian 3.3 tentang instrumen penelitian.



Gambar 3.2 Alur Kegiatan pada Tahapan Studi Perskriptif

3.1.3.1 Persiapan Data

Pada kegiatan ini prosesnya meliputi pencarian, pengumpulan, dan perbaikan *dataset* yang akan digunakan. Pengumpulan *dataset* dilakukan dengan mengunduh atau mengekstrak *dataset* yang sudah tersedia di internet. Sedangkan perbaikan *dataset* meliputi apabila terdapat data atau fitur yang mengakibatkan *dataset* tidak dapat digunakan. Penjelasan untuk *dataset* yang digunakan dijelaskan pada bagian 3.2.

3.1.3.2 *Data Cleaning*

Kegiatan pembersihan data atau *data cleaining* dilakukan pada *dataset* yang sudah disiapkan. Proses *data cleaning* dapat meliputi teknik-teknik pada umumnya seperti memilih kolom-kolom dari *dataset* yang relevan untuk pelatihan model, melakukan standarisasi data, dan normalisasi data. Kegiatan *data cleaning* ini bertujuan untuk meningkatkan kualitas data dan mengoptimalkan performa sebelum diproses oleh model *machine learning* (Fatima dkk., 2017).

Menurut Assudani dan Wankhede (2022), teknik-teknik *data cleaning* meliputi menangani data yang hilang dan penghapusan nilai ekstrem. Selain itu,

Fatima dkk. juga menyebutkan bahwa teknik memperbaiki data dan penyesuaian data dengan kasus dunia nyata memberikan konsistensi terhadap *dataset*.

Bertumpu dari penelitian sebelumnya, proses data *cleaning* pada penelitian ini akan melakukan beberapa kegiatan. Menghapus beberapa bagian dari *dataset* seperti data duplikat, kolom-kolom yang tidak relevan, data dengan nilai kosong, data harga rumah yang hanya menjual tanah, dan beberapa nilai ekstrem. Menghapus data yang hanya menjual tanah dilakukan untuk menjaga konsistensi *dataset* dan penelitian yang membahas tentang prediksi harga rumah. Dengan menghapus data-data yang terduplikat dan menghapus data dengan nilai kosong, hal ini dapat berguna untuk menjaga integritas data sehingga tidak menimbulkan bias terhadap suatu data (Fatima dkk., 2017). Mengurangi nilai ekstrem atau *outlier* juga dilakukan karena nilai ekstrem pada *dataset* dapat berpengaruh terhadap performa atau akurasi model menjadi lebih buruk (Tang dkk., 2022). Namun, bukan berarti *outlier* dihapus secara keseluruhan. Penghapusan keseluruhan *outlier* dapat mengakibatkan bias data terhadap perbedaan kelompok data (Karch, 2022).

Selain itu beberapa kegiatan transformasi data dilakukan juga pada proses data *cleaning* untuk penelitian ini seperti mengubah tipe data, *encoding categorical data*, dan normalisasi data. Mengubah tipe data dari suatu *feature* atau kolom dan *encoding categorical data* atau mengubah data kategori ke dalam bentuk *number* bertujuan agar data-data harga rumah dapat diproses oleh model-model prediksi menggunakan *machine learning* yang sudah dibuat. Lalu, normalisasi data dilakukan bertujuan untuk mengubah data ke dalam bentuk yang lebih seragam sehingga tidak ada data yang memiliki nilai yang mendominasi data lainnya. Selain itu, interpretasi nilai juga menjadi lebih konsisten dan meningkatkan performa analisis model (Firmansyah, 2024).

3.1.3.3 *Data Splitting*

Pemisahan data atau *data splitting* merupakan kegiatan yang membagikan data menjadi tiga bagian: data *Training,* data uji, dan data validasi.

Data *Training* adalah data yang digunakan oleh model sebagai bahan untuk mempelajari bagaimana data tersebut saling terhubung. Data *Training* ini juga digunakan oleh model *machine learning* sebagai acuan untuk kemampuan

memprediksi, sehingga data ini menjadi salah satu faktor penting yang mempengaruhi tingkat akurasi model.

Sedangkan data uji adalah data yang digunakan oleh model untuk mengevaluasi dirinya dalam kemampuan memprediksi. Dengan data ini, model *machine learning* dapat dinilai seberapa baik model belajar dari data *Training*.

Data validasi pada penelitian ini digunakan sebagai pengaturan parameter model dan melihat *overfitting* pada model. Hal ini sesuai dengan penjelasan pada penelitian yang dilakukan oleh Bilmes (2020) dengan melihat jarak antara hasil *Training* dan validasi. Untuk pengaturan parameter dilakukan pada model algoritma ANN dalam proses *hyperparameter tuning*.

Pada penelitian yang dilakukan oleh Rahayuningtyas dkk. (2021) dan Saiful dkk. (2021) rasio data yang digunakan adalah data *Training* dan data uji dengan jumlah 70:30 dan 80:20. Selain itu, penelitian lain yang dilakukan oleh Xu dan Zhang (2021) menggunakan rasio data *Training*, data uji, dan data validasi sebesar 80:10:10, 70:15:15, dan 60:20:20. Pembagian rasio data yang serupa dilakukan juga oleh Muneeb (2022) dengan menggunakan rasio data *Training*, data uji, dan data validasi sebesar 50:25:25. Meskipun begitu, pada penelitian ini akan dilakukan eksplorasi pada Rasio data *Training*, data uji, dan data validasi dengan menambahkan rasio pembagian data-datanya. Adapun rasio data yang akan digunakan pada penelitian ini dibagi menjadi empat: 80:10:10, 70:15:15, 60:20:20, dan 50:25:25. Setiap rasio data ini diuji pada algoritma ANN dan *Random Forest*. Hal ini dilakukan untuk menentukan rasio pembagian data mana yang menghasilkan performa terbaik pada model.

3.1.3.4 Arsitektur Model

Untuk menentukan parameter dan arsitektur yang tepat pada model ANN, teknik *hyperparameter tuning* digunakan. Hal ini untuk mengurangi *overfitting* pada model dan mengoptimasi tingkat performa model (Calugar dkk., 2022).

Model prediksi harga rumah dibuat dengan menggunakan algoritma ANN dan *Random Forest*. Selain itu, setiap model akan dibuat dengan implementasi rasio data *Training* dan data uji yang berbeda-beda sesuai tertera pada 3.1.3.3.

3.1.3.5 *Training* dan Pengujian Model

Setelah data sudah dipisahkan dan arsitektur model sudah dilakukan, tahap selanjutnya adalah mengimplementasikan data-data *Training* dan uji kepada model. Hasil *Training* dan pengujian model akan dibandingkan dengan model lainnya untuk dievaluasi performa setiap modelnya. Selain itu, analisis terhadap hasil validasi dilakukan untuk melihat seberapa besar *overfitting* terjadi pada model.

3.1.4 Studi Deskriptif 2

Pada tahapan ini, hasil perhitungan dari model algoritma yang sudah dibuat dihitung ke dalam bentuk berdasarkan ukuran metrik evaluasi yang sudah ditentukan. Hasil dari model-model yang sudah dilatih dan diuji akan dievaluasi dengan membandingkan hasil prediksi dari setiap model. Proses evaluasi perbandingan ini dilakukan untuk melihat dari performa setiap model. Kesimpulan dan hipotesis dibuat berdasarkan proses evaluasi perbandingan. Hal ini untuk memberikan gambaran mengenai hasil penelitian yang sudah dilakukan.

3.2 *Dataset* Harga Rumah Di Kota Bandung

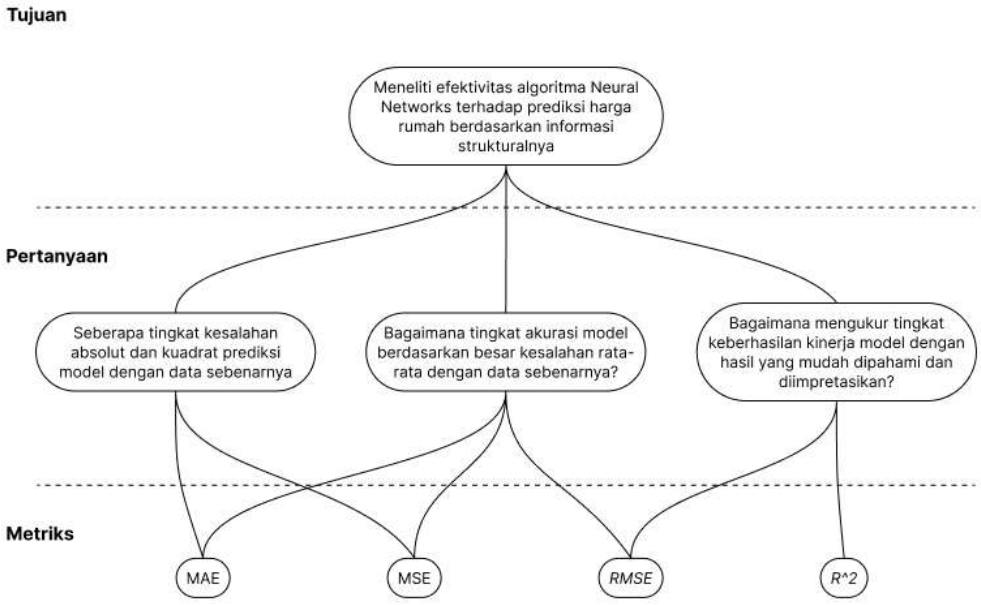
Proses persiapan data menggunakan *Dataset* yang didapat dari laman Kaggle yang berjudul "Data Harga Rumah di Kota Bandung". *Dataset* ini dikumpulkan dari *website* rumah123.com pada bulan Maret tahun 2024 melalui proses *web scrapping* oleh Al Faaath (2024). *Dataset* ini berisikan data-data harga rumah di daerah Kota Bandung, Jawa Barat beserta dengan data nama rumah, cicilan, apakah halaman masuk ke dalam *premier* atau *featured* website, jenis, harga, lokasi berdasarkan kecamatannya, dan karakteristik bentuk struktural rumahnya. *Dataset* yang tersedia memiliki format fail .csv dengan jumlah data yang tersedia sebanyak 7.611 baris dengan 11 kolom.

3.3 Instrumen Penelitian

Terdapat empat instrumen penelitian yang digunakan dalam penelitian ini, yaitu *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE), dan *R Square* (R^2). Metrik RMSE baik digunakan dalam model yang bertujuan untuk memprediksi (Manasa dkk., 2020). Hodson (2022) dan

Chai dan Draxler (2014) menyarankan bahwa penggunaan metrik RMSE sebaiknya digunakan bersamaan dengan metrik MAE.

Pertanyaan yang terhubung dengan instrumen penelitian metrik yang digunakan tertera pada diagram *Goal Question Metrics* (GQM) di Gambar 3.3.



Gambar 3.3 Diagram GQM dari instrumen penelitian yang digunakan

3.3.1 *Mean Absolut Error* (MAE)

MAE didefinisikan sebagai metode evaluasi model dengan menghitung rata-rata atau *mean* dari selisih absolut antara data sebenarnya dengan data hasil prediksi. Metrik evaluasi ini digunakan untuk memberikan nilai evaluasi seberapa besar model meleset dalam memprediksi dengan rata-rata nilai aslinya (Hodson, 2022). Untuk kasus penelitian ini, MAE mengukur selisih antara data harga rumah sebenarnya dengan data harga rumah hasil prediksi model. Semakin kecil MAE maka semakin baik performa model tersebut (Sharma dkk., 2021). Adapun formula perhitungan metrik MAE disebutkan dalam Formula 3.1.

$$MAE = \frac{\Sigma|Y' - Y|}{n}$$

Formula 3.1 Rumus *Mean Absolut Error* (MAE) (Rahayuningtyas dkk., 2021)

Dimana:

Y' = nilai prediksi

Y = nilai sebenarnya

n = jumlah data

3.3.2 *Mean Squared Error* (MSE)

MSE adalah metrik evaluasi model yang menghitung rata-rata dari selisih kuadrat antara nilai asli atau nilai harga rumah sebenarnya dengan nilai harga rumah dari prediksi model, atau untuk penggambaran rumusnya ada pada Formula 3.2. Metrik ini cenderung sensitif terhadap nilai ekstrem atau *outlier* (Plevris dkk., 2022). Maka dari itu, MSE dapat digunakan untuk memberikan informasi apakah ada data ekstrem atau *outlier* yang sangat berpengaruh terhadap performa modelnya.

$$MSE = \frac{\Sigma|Y' - Y|^2}{n}$$

Formula 3.2 Rumus *Mean Squared Error* (MSE) (Rahayuningtyas dkk., 2021)

3.3.3 *Root Mean Squared Error* (RMSE)

Metrik RMSE mengukur performa model dengan menghitung kuadratik dari MSE. Menurut Chai dan Draxler (2014), RMSE baik digunakan untuk melihat apakah model memiliki isu dalam performanya dengan memberikan bobot lebih pada *error* atau kesalahan yang lebih besar. Selain itu, RMSE juga menghindari nilai yang absolut yang dimana lebih sering dihindari dalam banyak kalkulasi matematika, terutama apabila datanya cenderung terdistribusi *gaussian*. Adapun formula perhitungan metrik RMSE tercantum pada Formula 3.3.

$$RMSE = \sqrt{\frac{\Sigma|Y' - Y|^2}{n}}$$

Formula 3.3 Rumus *Root Mean Squared Error* (RMSE) (Chai dan Draxler, 2014)

3.3.4 *R Square* (R^2)

Metrik koefisien Determinasi atau *R square* merupakan salah satu metrik yang sering digunakan dalam mengukur performa dari suatu model regresi. Metrik ini digunakan *R Square* mengukur rasio variabilitas atau persebaran yang dapat dipahami oleh model, yang artinya metrik ini menghitung proporsi varian antara data variabel sebenarnya dengan variabel prediksi (Plevris dkk., 2022).

Menurut Chicco dkk. (2021), metrik ini memiliki nilai batas atas yang jelas, berbeda dengan MAE, MSE, dan RMSE yang tidak memiliki batas atas sehingga memungkinkan hasil metrik-metrik tersebut bernilai positif tak hingga. Hal ini menjadikan metrik *R square* lebih informatif dan mudah dibaca.

Metrik *R square* bernilai antara 0 dan 1. Berbeda dengan metrik-metrik sebelumnya, nilai *R squared* yang mendekati 1 menunjukkan bahwa model dapat menjelaskan sebagian besar variabilitas data. Adapun rumus *R square* dijelaskan pada Formula 3.4.

$$R^2 = 1 - \frac{\sum(Y - Y')^2}{\sum(Y - \bar{Y})^2}$$

Formula 3.4 Rumus R square yang Digunakan Untuk Model Regresi (Plevris dkk., 2022)

Dimana:

$\bar{Y}$ = rata-rata nilai sebenarnya

3.4 Alat dan Bahan Penelitian

Untuk alat penelitian yang digunakan adalah perangkat laptop dan bahasa pemrograman Python dengan versi 3.10. Bahasa pemrograman Python digunakan sebagai bahasa pemrosesan data pada model algoritma dan dilengkapi dengan banyak *library* yang memudahkan dalam pemrosesan data dan perancangan model *machine learning*. Penggunaan Alat yang digunakan tertera pada Tabel 3.1.

Tabel 3.1

Spesifikasi Perangkat yang Digunakan

Komponen	Spesifikasi

Sistem operasi	Windows 11
Random Access Memory (RAM)	DDR4 8GB (3200 MHz)
Processor	Intel Core i5-11400H 2.70GHz
Perangkat penyimpanan	*Solid State Drive* (SSD)
Graphic Processing Unit (GPU)	GeForce RTX3050 Max-Q, GDDR6 4GB

Adapun perangkat lunak yang digunakan beserta *library*-nya dicantumkan pada Tabel 3.2.

Tabel 3.2

Perangkat lunak dan *library* yang digunakan

Nama	Jenis	Deskripsi
Browser Microsoft Edge	Aplikasi	Digunakan untuk perancangan model prediksi harga rumah dan kajian literatur
Google Colab	*Platform*	Digunakan untuk perancangan model prediksi harga rumah
matplotlib.pyplot	*Library*	Sebagai visualisasi data
Pandas	*Library*	Sebagai alat untuk memanipulasi data
Tensorflow	*Library*	Untuk perancangan model *machine learning*
Seaborn	*Library*	Untuk visualisasi data, terutama korelasi antar *feature* dataset
Numpy	*Library*	Untuk manipulasi data metrik dan *array*
SkLearn	*Library*	Sebagai alat untuk memanipulasi data dan pembuatan model *machine learning*