

Kelompok 7

11 November 2025

**Daftar Isi**

<b>1 Judul Proyek</b>	<b>1</b>
<b>2 Latar Belakang Masalah</b>	<b>2</b>
<b>3 Rumusan Masalah dan Tujuan Proyek</b>	<b>2</b>
3.1 Rumusan Masalah . . . . .	2
3.2 Tujuan Proyek . . . . .	2
<b>4 Sumber dan Karakteristik Data</b>	<b>3</b>
4.1 Sumber Data . . . . .	3
4.2 Karakteristik Data . . . . .	3
<b>5 Metodologi Data Mining</b>	<b>3</b>
<b>6 Teknik dan Algoritma yang Digunakan</b>	<b>4</b>
6.1 Fungsi Minor . . . . .	4
6.2 Fungsi Mayor . . . . .	4
<b>7 Rencana Implementasi dan Alat yang Digunakan</b>	<b>5</b>
7.1 Alat & Library . . . . .	5
<b>8 Hasil yang Diharapkan dan Dampak</b>	<b>5</b>
8.1 Hasil . . . . .	5
8.2 Dampak . . . . .	5

**Daftar Gambar****1 Judul Proyek**

Prediksi Harga Rumah Menggunakan Regresi Linear pada Dataset Boston Housing

## 2 Latar Belakang Masalah

Pasar properti merupakan salah satu sektor ekonomi yang sangat kompleks dan dinamis karena melibatkan berbagai faktor sosial, ekonomi, lingkungan, dan demografi yang saling mempengaruhi. Harga rumah tidak hanya bergantung pada atribut fisik rumah seperti luas bangunan dan lokasi geografis, tetapi juga sangat dipengaruhi oleh karakteristik sosial-ekonomi masyarakat di sekitarnya. Faktor sosial seperti tingkat kejahatan, kualitas lingkungan, fasilitas pendidikan, dan komposisi demografis penduduk telah terbukti memiliki hubungan signifikan terhadap nilai properti (Harrison & Rubinfeld, 1978). Studi lain juga menegaskan pentingnya variabel lingkungan dan sosial-ekonomi dalam analisis harga rumah di kawasan perkotaan (Hidayah et al.) (Chen 2025)

Boston Housing Dataset merupakan salah satu dataset yang paling banyak dipakai untuk analisis harga rumah karena mengandung informasi detail tentang faktor-faktor yang mempengaruhi harga rumah di wilayah Boston (Puneeth et al.). Dataset ini meliputi 14 variabel, seperti tingkat kejahatan di lingkungan sekitar, kualitas sekolah, jarak ke pusat kerja, kualitas udara, serta demografi penduduk, khususnya proporsi penduduk berkulit hitam yang dapat memengaruhi harga rumah (Harrison & Rubinfeld, 1978). Dataset ini menjadi acuan penting dalam penelitian regresi dan machine learning untuk membangun model prediksi harga rumah dengan pendekatan data mining (Hartarti et al. 2025).

Metode regresi linear dipilih dalam penelitian ini karena kemampuannya dalam mengukur dan mengestimasi hubungan linier antara berbagai faktor sosial-ekonomi dan harga rumah. Dengan regresi linear, prediksi harga rumah dapat dihasilkan secara kuantitatif sehingga dapat mendukung pengambilan keputusan pengguna, pengembang properti, dan pembuat kebijakan. Selain itu, penelitian komparatif model regresi linear dengan metode machine learning lain juga menunjukkan bahwa regresi linear memberikan interpretabilitas yang jelas dalam memahami pengaruh setiap variabel independen terhadap harga rumah.

Tujuan utama dari penelitian ini adalah mengidentifikasi dan mengukur faktor-faktor kunci yang berkontribusi pada nilai properti menggunakan model regresi linear pada Boston Housing Dataset. Dengan hasil analisis yang komprehensif, penelitian bertujuan untuk memberikan wawasan yang dapat membantu pihak terkait dalam menganalisis dinamika pasar properti serta merumuskan strategi investasi dan regulasi hunian yang lebih efektif (Rianti et al. 2023). Penelitian serupa telah berhasil mengaplikasikan regresi untuk memodelkan harga rumah dan memperlihatkan tren kenaikan harga yang dipengaruhi oleh perubahan faktor sosial-ekonomi dan lokasi.

## 3 Rumusan Masalah dan Tujuan Proyek

### 3.1 Rumusan Masalah

1. Faktor apa saja yang paling memengaruhi harga median rumah di wilayah Boston?
2. Bagaimana pengaruh variabel sosial seperti proporsi penduduk berkulit hitam (black) terhadap harga rumah?
3. Seberapa besar pengaruh masing-masing variabel terhadap harga rumah?
4. Seberapa akurat model regresi linear dalam memprediksi harga rumah berdasarkan variabel-variabel tersebut?

### 3.2 Tujuan Proyek

- Menganalisis hubungan antara variabel-variabel sosial, ekonomi, dan fisik terhadap harga rumah.

- Menilai sejauh mana proporsi penduduk berkulit hitam berpengaruh terhadap nilai properti.
- Membangun model regresi linear berganda untuk memprediksi harga median rumah.
- Mengevaluasi akurasi dan reliabilitas model dengan menggunakan metrik statistik.

## 4 Sumber dan Karakteristik Data

### 4.1 Sumber Data

Dataset Boston berasal dari paket MASS dalam R ([people.sc.fsu.edu](http://people.sc.fsu.edu)).

### 4.2 Karakteristik Data

- Jumlah observasi: 506 rumah
- Variabel: 14 (13 prediktor dan 1 target medv)
- Variabel kunci:
  - rm (rata-rata jumlah kamar per rumah)
  - lstat (persentase populasi berstatus sosial ekonomi rendah)
  - nox (polusi udara, dalam ppm)
  - dis (jarak properti ke lima pusat bisnis utama Boston)
  - black (1000 × (Bk – 0.63), proporsi penduduk berkulit hitam)
  - crim (tingkat kejahatan per kapita)
  - medv (harga median rumah dalam \$1000)
- Potensi tantangan: keberadaan multikolinearitas antar variabel sosial dan ekonomi (Zhang, 2023).

## 5 Metodologi Data Mining

Metodologi penelitian menggunakan model CRISP-DM (Cross Industry Standard Process for Data Mining) yang terdiri dari enam tahapan utama: pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan deployment (Rianti et al. 2023). Model ini dipilih karena sudah menjadi standar de facto dalam proyek data mining dan data science, serta memberikan kerangka kerja yang sistematis dan terstruktur untuk mengelola proses penambangan data dari awal hingga akhir (Schröer et al. 2021). Setiap tahapan dalam CRISP-DM dapat disesuaikan dengan kebutuhan proyek, sehingga fleksibel dan dapat diterapkan secara luas di berbagai domain penelitian (Rianti et al. 2023).

### 1. Business Understanding

Tahapan ini fokus untuk memahami konteks bisnis atau masalah yang akan dipecahkan dengan data mining. Pada tahap ini, penelitian menentukan kebutuhan dan tujuan bisnis, termasuk faktor sosial dan ekonomi yang berpengaruh pada harga properti, sehingga permasalahan dapat dirumuskan dengan jelas untuk proses analisis data selanjutnya. Tahap ini sangat penting untuk memastikan bahwa hasil data mining dapat memberikan solusi yang relevan dan bermanfaat dari sudut pandang bisnis (Ruswanti et al. 2024)

### 2. Data Understanding

Tahap ini melibatkan pengumpulan data, analisis statistik deskriptif, dan visualisasi data untuk mengetahui pola dan hubungan antar variabel. Contohnya adalah penggunaan corplot dan ggplot untuk

memvisualisasikan korelasi maupun sebaran variabel dalam dataset. Selain itu, tahap ini juga mengevaluasi kualitas data sebelum dilakukan pemodelan lebih lanjut, sehingga kesalahan dan inkonsistensi dapat diidentifikasi (Rianti et al. 2023)

### 3. Data Preparation

Normalisasi variabel numerik. Pembersihan outlier pada medv dan lstat. Seleksi fitur menggunakan korelasi dan uji signifikansi.

### 4. Modeling

Pada tahap modeling, dilakukan pembuatan model regresi linear sederhana dan berganda untuk memprediksi nilai median harga rumah (medv) berdasarkan variabel independen pada Boston Housing Dataset (Usman et al. 2025). Model yang dibuat menggunakan rumus regresi linear multiple, yaitu medv sebagai variabel dependen diprediksi oleh variabel-variabel independen seperti jumlah kamar rata-rata per rumah (rm), persentase populasi dengan status sosial rendah (lstat), jarak ke pusat kerja (dis), tingkat kriminalitas per kapita (crim), serta variabel sosial-black (black) yang merepresentasikan proporsi penduduk berkulit hitam di wilayah tersebut (geeksforgeeks.org 2025). Fungsi lm() di R digunakan untuk membuat model ini dengan sintaks:

```
model <- lm(medv ~ rm + lstat + nos + dis + crim + black, data = Boston) summary(model)
```

Model ini dievaluasi menggunakan fungsi summary() untuk memperoleh ringkasan statistik termasuk koefisien regresi, nilai p, dan koefisien determinasi (R-squared) yang menunjukkan seberapa baik variabel independen dapat menjelaskan variasi harga rumah. Pendekatan multiple linear regression ini memberikan kemampuan untuk mengestimasi pengaruh setiap faktor secara linier dan simultan terhadap harga rumah, sehingga model dapat dijadikan dasar analisis kuantitatif yang valid dalam konteks sosial-ekonomi dan perencanaan properti (Hartarti et al. 2025). Selain itu, metode ini banyak digunakan dalam studi prediksi harga properti karena interpretabilitas dan hasil yang dapat diandalkan untuk pengambilan keputusan (Salsabila 2023) \* Fungsi mayor: Regresi Linear.

### 5. Evaluation

Analisis goodness of fit: R<sup>2</sup>, adjusted R<sup>2</sup>, p-value, F-test.

### 6. Deployment (opsional)

Visualisasi hasil dan interpretasi variabel paling berpengaruh terhadap harga.

## 6 Teknik dan Algoritma yang Digunakan

### 6.1 Fungsi Minor

- Pembersihan data (na.omit, boxplot.stats)
- Normalisasi dan transformasi log jika diperlukan
- Seleksi fitur berdasarkan korelasi tinggi terhadap medv

### 6.2 Fungsi Mayor

- Regresi Linear Berganda untuk prediksi harga rumah
- Interpretasi koefisien model sebagai dasar kebijakan sosial-ekonomi

Contoh penerapan fungsi mayor:

```
library(MASS) data(Boston) model <- lm(medv ~ rm + lstat + nos + dis + crim + black, data = Boston) summary(model)
```

## 7 Rencana Implementasi dan Alat yang Digunakan

### 7.1 Alat & Library

- R & RStudio
- Paket: MASS, ggplot2, dplyr, corrplot, caret, broom

## 8 Hasil yang Diharapkan dan Dampak

### 8.1 Hasil

- Model prediksi harga rumah berdasarkan faktor sosial, ekonomi, dan lingkungan.
- Penilaian empiris terhadap pengaruh proporsi penduduk berkulit hitam terhadap harga properti.

### 8.2 Dampak

- Memberikan wawasan terhadap bias sosial dan diskriminasi pasar properti yang masih terjadi di kota besar.
- Menjadi dasar pengambilan keputusan yang lebih adil dan berbasis data dalam kebijakan perumahan.

### Daftar Pustaka

Chen Z. 2025. *House Price Prediction in Boston Based on BP Neural Network*. *Highlights Sci. Eng. Technol.* 124:152–156. doi:10.54097/2vq1wq81.

geeksforgeeks.org. 2025. *Multiple linear regression analysis of Boston Housing Dataset using R*.

Hartarti ID, Septiyani IA, Gultom DA, Hendrian Y, Kinanti SL. 2025. *Prediksi Harga Rumah di Boston Dengan Model Regresi Linear Menggunakan Python*. *RIGGS J. Artif. Intell. Digit. Bus.* 4(2):4250–4256. doi:10.31004/riggs.v4i2.1210.

Hidayah F, Angesti SJ, Widyastuti YP. *Prediksi Harga Rumah di Boston Menggunakan Metode Linear Regression, SVR, Decision Tree dan Random Forest Regression*.

Puneeth S, Quadri A, Sahithi M, Arbas M, Jyothi PS. *PREDICTING HOME PRICES: A BEGINNER'S JOURNEY WITH REGRESSION ANALYSIS USING THE BOSTON HOUSING DATASET*.

Rianti A, Majid NWA, Fauzi A. 2023. *CRISP-DM: Metodologi Proyek Data Science*.

Ruswanti D, Susilo D, Riani R. 2024. *Implementasi CRISP-DM pada Data Mining untuk Melakukan Prediksi Pendapatan dengan Algoritma C.45*. *Go Infotech J. Ilm. STMIK AUB.* 30(1):111–121. doi:10.36309/goi.v30i1.266.

Salsabila S. 2023. *Dataset Boston Housing dengan R*.

*Schröer C, Kruse F, Gómez JM. 2021. A Systematic Literature Review on Applying CRISP-DM Process Model. Procedia Comput. Sci. 181:526–534.doi:10.1016/j.procs.2021.01.199.*

*Usman N, Thamrin M, Faisal M. 2025. COMPARISON OF THE PERFORMANCE OF REGRESSION-SPECIFIC AND MULTI-PURPOSE ALGORITHMS. 4(8).*