

Universitas Katolik Santo Thomas

Repositori Unika Santo Thomas

<http://eprints.ust.ac.id>

Fakultas Ilmu Komputer (FIKOM)
Program Studi Teknik Informatika

in the Undergraduate Papers

Pasaribu, Ishak Eben Haezer

2025

Analisis Prediksi Harga Rumah Di Kota Medan Berdasarkan Spesifikasi Dengan Model Regresi Linear Dan Random Forest

<http://eprints.ust.ac.id/id/eprint/1048>

Downloaded from Repositori Institusi UST, Universitas Katolik Santo Thomas

**ANALISIS PREDIKSI HARGA RUMAH DI KOTA MEDAN
BERDASARKAN SPESIFIKASI DENGAN MODEL
REGRESI LINEAR DAN RANDOM FOREST**

SKRIPSI

**Diajukan Untuk Melengkapi dan Memenuhi Persyaratan
Memperoleh Gelar Sarjana Komputer Pada
Program Studi Teknik Informatika**

Oleh:

ISHAK EBEN HAEZER PASARIBU

200840079



**FAKULTAS ILMU KOMPUTER
UNIVERSITAS KATOLIK SANTO THOMAS
SUMATERA UTARA
MEDAN
2025**

PERNYATAAN PENULIS

Sebagai sivitas akademik Universitas Katolik Santo Thomas, saya yang bertanda tangan di bawah ini:

Nama : Ishak Eben Haezer Pasaribu
NPM : 200840079
Program Studi : Teknik Informatika
Fakultas : Ilmu Komputer
Jenis Karya : Skripsi

1. Menyatakan dan bertanggungjawab dengan sebenarnya bahwa Skripsi ini adalah hasil karya saya sendiri, kecuali cuplikan dan ringkasan yang masing-masing telah saya jelaskan sumbernya. Jika pada waktu selanjutnya ada pihak lain yang mengklaim bahwa Skripsi ini sebagai karyanya, yang disertai dengan bukti-bukti yang cukup, maka saya bersedia untuk dibatalkan gelar Sarjana Komputer saya beserta segala hak dan kewajiban yang melekat pada gelar tersebut.
2. Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Katolik Santo Thomas Hak Bebas Royalti Nonesklusif (Non Exclusive Royalty-Free Right) atas karya ilmiah saya tersebut, beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Nonesklusif ini, Universitas Katolik Santo Thomas berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (database), merawat, dan mempublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis / pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Universitas Katolik Santo Thomas

Pada Tanggal : 25 September 2025

Yang Menyatakan

Ishak Eben Haezer Pasaribu
Penulis

LEMBAR PENGESAHAN SKRIPSI

NAMA : ISHAK EBEN HAEZER PASARIBU
NPM : 200840079
PROGRAM STUDI : TEKNIK INFORMATIKA
FAKULTAS : ILMU KOMPUTER
JUDUL : ANALISIS PREDIKSI HARGA RUMAH DI
KOTA MEDAN BERDASARKAN SPESIFIKASI
DENGAN MODEL REGRESI LINEAR DAN
RANDOM FOREST



SKRIPSI

ANALISIS PREDIKSI HARGA RUMAH DI KOTA MEDAN BERDASARKAN SPESIFIKASI DENGAN MODEL REGRESI LINEAR DAN RANDOM FOREST

Dipersiapkan dan Diajukan oleh :

ISHAK EDEN HAEZER PASARIBU

200840079

Telah dipertahankan di depan Dewan Penguji

Pada Tanggal 26 September 2025

UNIVERSITAS KATOLIK SANTO THOMAS

Susunan Dewan Penguji

OMNIBUS OMNIA

Pembimbing Utama/Penguji I

Penguji II

UNIKA SANTO THOMAS

Prof. Dr. Zakarias Situmorang, M.T., MCE
NIDN. 0118127801

Nayel Singian, S.Kom., M.Kom
NIDN. 0129119401

Penguji III

Desinta Purba, S.T., M.Kom
NIDN. 0111067801

ABSTRAK

Kesulitan masyarakat dalam menentukan harga rumah di Kota Medan akibat variasi informasi mendorong kebutuhan akan model prediksi yang akurat. Penelitian ini mengisi kesenjangan dengan secara komprehensif membandingkan model Regresi Linear dan Random Forest menggunakan data hasil web scraping dari situs properti daring Rumah123.com sebanyak 6748 baris data. Tujuan penelitian ini adalah menganalisis, melatih, dan mengevaluasi kedua model tersebut untuk menentukan mana yang paling optimal dalam memprediksi harga rumah di Kota Medan. Metode penelitian menggunakan pendekatan komparatif, di mana dua model dilatih menggunakan library Scikit-Learn dan dievaluasi performanya berdasarkan metrik Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), dan R^2 . Hasil menunjukkan bahwa model Random Forest Regressor memiliki performa terbaik dengan nilai R^2 sebesar 0.86 pada data latih dan 0.77 pada data uji. Analisis fitur dari model terbaik mengidentifikasi luas tanah (41%) dan luas bangunan (36%) sebagai faktor yang paling berpengaruh terhadap harga. Disimpulkan bahwa Random Forest merupakan model yang lebih unggul untuk studi kasus ini. Disarankan bagi penelitian selanjutnya untuk mengeksplorasi model lain seperti yang berbasis ensemble dan memperkaya dataset untuk meningkatkan generalisasi model.

Kata Kunci: Prediksi Harga Rumah, Regresi Linear, Random Forest, Web Scraping, Machine Learning, Kota Medan

ABSTRACT

Facilitating the public in determining house prices in Medan City due to the variety of information drives the need for accurate prediction models. This study fills the whole by comprehensively comparing the Linear Regression and Random Forest models using web scraping data from the bold property site Rumah123.com totaling 6748 rows of data. The purpose of this study is to analyze, train, and emit both models to determine which is the most optimal in predicting house prices in Medan City. The research method uses a comparative approach, where two models are drilled using the Scikit-Learn library and their performance is evaluated based on the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R2 metrics. The results show that the Random Forest Regressor model has the best performance with an R2 value of 0.86 on the training data and 0.77 on the test data. Feature analysis of the best model in identifying land area (41%) and building area (36%) as the most influential factors on prices. It is concluded that Random Forest is a superior model for this case study. It is recommended for further research to explore other models such as ensemble-based ones and enrich the dataset to improve model generalization.

Keywords: *House Price Prediction, Linear Regression, Random Forest, Web Scraping, Machine Learning, Medan City*

KATA PENGANTAR

Puji syukur penulis panjatkan kepada Tuhan Yang Maha Esa, atas berkat rahmat dan karunia-Nya akhirnya penulis dapat menyelesaikan skripsi.

Skripsi ini berjudul Analisis Prediksi Harga Rumah Di Kota Medan Berdasarkan Spesifikasi Dengan Model Regresi Linear Dan Random Forest yang diajukan untuk memperoleh gelar Sarjana Dalam penyusunan skripsi ini, penulis banyak memperoleh support dari berbagai pihak, oleh karena itu penulis menyampaikan penghargaan dan ucapan terima kasih yang sebesar-besarnya kepada :

1. Kepada kedua orangtua penulis, D. Pasaribu dan R. Br. Sitohang yang tidak pernah lelah memberikan nasihat, dukungan, semangat, dan selalu mendoakan penulis.
2. Kepada Bapak Prof. Dr. Maidin Gultom, SH, M.Hum selaku Rektor Universitas Katolik Santo Thomas Sumatera Utara.
3. Kepada Ibu Desinta Purba, ST, M.Kom selaku dekan fakultas ilmu komputer Universitas Katolik Santo Thomas.
4. Kepada Ibu Masdiana Sagala, S.Kom. M.Kom selaku ketua program studi teknik informatika.
5. Kepada Prof. Dr. Zakarias Situmorang, M.T., MCE selaku dosen pembimbing skripsi yang telah memberikan arahan, bimbingan, dan dukungan selama proses penyusunan laporan ini dengan sabar.
6. Rekan-rekan penulis yang telah banyak memberikan bantuan ikut berperan dalam memperlancar penelitian dan penulisan skripsi ini.

Penulis menyadari bahwa Skripsi ini masih terdapat kekurangan, untuk itu penulis sangat mengharapkan saran-saran dan kritikan yang dapat membangun dan memperbaiki skripsi ini. Semoga skripsi ini bermanfaat bagi setiap pembaca, khususnya bagi mahasiswa Teknik Informatika.

Akhir kata penulis mengucapkan terimakasih.

Medan, 25 September 2025

Penulis,

Ishak Eben Haezer Pasaribu

DAFTAR ISI

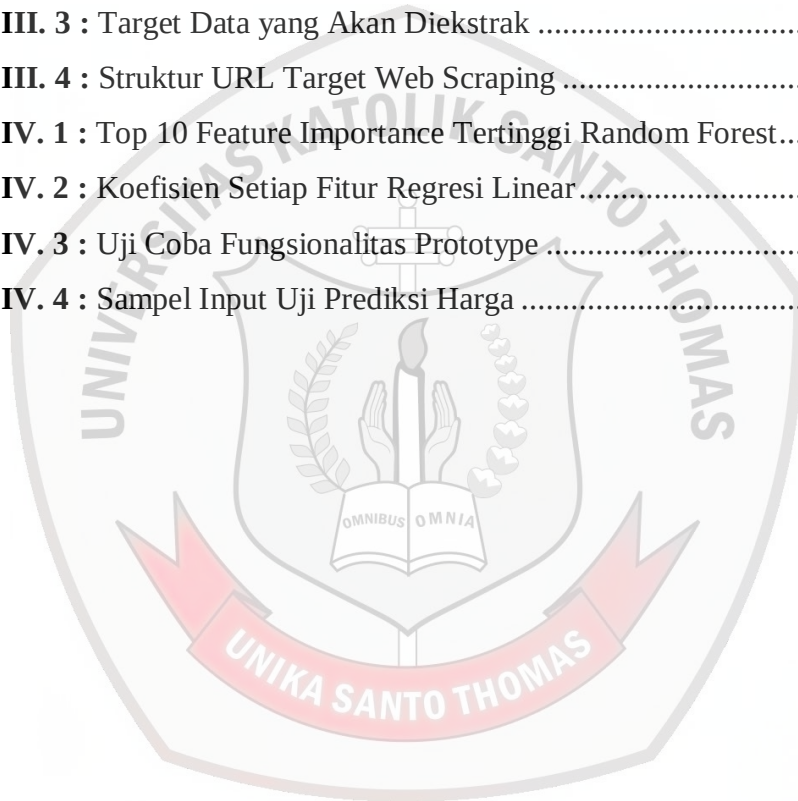
PERNYATAAN PENULIS.....	i
LEMBAR PENGESAHAN SKRIPSI.....	Error! Bookmark not defined.
ABSTRAK.....	ii
KATA PENGANTAR.....	vi
DAFTAR ISI	vii
DAFTAR TABEL	x
DAFTAR GAMBAR.....	xi
BAB I PENDAHULUAN	1
I.1. Latar Belakang Masalah.....	1
I.2. Rumusan Masalah.....	3
I.3 Batasan Masalah	3
I.4 Maksud dan Tujuan.....	4
I.5 Manfaat Penelitian	4
I.5.1. Manfaat Teoritis	4
I.5.2. Manfaat Praktis	4
I.6. Metodologi Penelitian.....	4
I.7. Sistematika Penulisan.....	5
BAB II LANDASAN TEORI.....	7
II.1. Prediksi.....	7
II.1.1. Prediksi Nilai Kontinu (Regresi).....	7
II.1.2. Prediksi Kategori (Klasifikasi)	9
II.2. Web Scraping	11
II.3. Jenis-Jenis Data	12
II.3.1. Data Kategorikal	12
II.3.2. Data Numerik.....	13
II.4. One Hot Encoding	14
II.5. Analisis Statistik	14
II.6. Regresi Linear	15
II.6.1. Regresi Linear Sederhana	16

II.6.2. Regresi Linear Berganda	17
II.6.3. Model Regresi Linear	18
II.6.4. Asumsi-Asumsi dalam Regresi Linear	19
II.6.5. Interpretasi Koefisien Regresi Linear.....	21
II.6.6. Interpretasi Hubungan antara Variabel.....	22
II.6.7. Kelebihan dan Kekurangan Regresi Linear	23
II.6.8. Perhitungan Regresi Linear	24
II.7. Random Forest.....	28
II.8. Python	32
II.9. Machine Learning.....	33
II.10. Jupyter Notebook.....	34
II.11. Unified Model Language.....	35
II.11.1. Use Case Diagram.....	36
II.11.2. Activity Diagram.....	37
BAB III ANALISIS DAN PERANCANGAN	39
III.1. Analisis Masalah	39
III.1.1. Studi Kasus.....	39
III.1.2. Permasalahan.....	39
III.3. Solusi	39
III.2. Kebutuhan Perangkat Lunak	40
III.3. Kebutuhan Perangkat Keras.....	40
III.4. Tahapan Penelitian	40
III.4.1. Web Scraping	41
III.4.2. Regression	44
III.4.3. Inference.....	48
III.4.4. Perancangan Prototype.....	49
BAB IV IMPLEMENTASI SISTEM	54
IV.1. Web Scraping	54
IV.2. Regression.....	56
IV.2.1. Data Understanding	56
IV.2.2. Penambahan Fitur	60

IV.2.3. Data Cleaning	61
IV.2.3. Exploratory Data Analysis	67
IV.2.4. Data Preparation	79
IV.2.5. Modelling	82
IV.2.6. Evaluation	84
IV.2.7. Export Model.....	98
IV.3. Inference	99
IV.3.1. Uji Coba Fungsionalitas.....	100
IV.3.2. Uji Coba Prediksi Model.....	102
BAB V PENUTUP.....	103
V.1. Kesimpulan	103
V.2. Saran	104
DAFTAR PUSTAKA	105
LAMPIRAN	108
Surat Pengangkatan Dosen Pembimbing Skripsi	108
Kartu Bimbingan Skripsi.....	109
Kartu Bimbingan Tugas Akhir.....	110

DAFTAR TABEL

Tabel II. 1 : Confusion Matrix	10
Tabel II. 2 : Sampel Data Iklan Rumah	24
Tabel II. 3 : Use Case Diagram	36
Tabel II. 4 : Activity Diagram.....	38
Tabel III. 1 : Kebutuhan Perangkat Lunak	40
Tabel III. 2 : Kebutuhan Perangkat Keras	40
Tabel III. 3 : Target Data yang Akan Diekstrak	43
Tabel III. 4 : Struktur URL Target Web Scraping	44
Tabel IV. 1 : Top 10 Feature Importance Tertinggi Random Forest.....	92
Tabel IV. 2 : Koefisien Setiap Fitur Regresi Linear	94
Tabel IV. 3 : Uji Coba Fungsionalitas Prototype	101
Tabel IV. 4 : Sampel Input Uji Prediksi Harga	102



DAFTAR GAMBAR

Gambar I.1 : List Iklan Rumah.....	2
Gambar II.1 :. Diagram Regresi Linear	16
Gambar II.2 : Hubungan Positif Linear	22
Gambar II.3 : Hubungan Negatif Linear	23
Gambar II.4 : Tidak Ada Hubungan	23
Gambar II.5 : Ilustrasi Random Forest Regression	29
Gambar II.6 : UI Jupyter Notebook.....	35
Gambar III.1 : Tahapan Penelitian	41
Gambar III.2 : Tahapan Webscraping	42
Gambar III.3 : Beranda Website	42
Gambar III.4 : Filter Rumah di Kota Medan	43
Gambar III.5 : Tahap Regression	45
Gambar III.6 : Use Case Diagram Prototype	49
Gambar III.7 : Activity Diagram Prediksi Harga	50
Gambar III.8 : Activity Diagram Melihat Model.....	51
Gambar III.9 : Rancangan UI – Tab Prediksi Harga	52
Gambar III.10 : Rancangan UI - Tab Informasi Model.....	52
Gambar IV.1 : Target Data yang Diekstrak	55
Gambar IV.2 : File CSV Hasil Web Scraping.....	56
Gambar IV.3 : Informasi Dataset.....	57
Gambar IV.4 : Statistik Deskriptif Dataset	58
Gambar IV.5 : Missing Values pada Dataset	60
Gambar IV.6 : Missing Values Setelah Penanganan	62
Gambar IV.7 : Boxplot Fitur Numerik Sebelum Penanganan Outlier	63
Gambar IV.8 : Jumlah Iklan Rumah Per Kecamatan di Kota Medan.....	64
Gambar IV.9 : Boxplot Setelah Dilakukan Penanganan Outliers	65
Gambar IV.10 : Informasi Dataset Setelah Cleaning.....	66
Gambar IV. 11 : Statistik Deskriptif Setelah Cleaning	66
Gambar IV. 12 : Distribusi Fitur Numerik	67

Gambar IV. 13 : Boxplot Fitur Numerik	68
Gambar IV.14 : Heatmap Jumlah Properti per Kecamatan.....	71
Gambar IV.15 : Scatterplot Harga Rumah vs. Luas Bangunan.....	72
Gambar IV.16 : Scatterplot Harga Rumah vs. Luas Tanah.....	73
Gambar IV.17 : Barchart Rata-Rata Harga Rumah per Kecamatan	74
Gambar IV.18 : Boxplot Harga Rumah per Kecamatan	75
Gambar IV.19 : Perbandingan Rata-Rata vs. Median Harga Rumah per Kecamatan.....	75
Gambar IV.20 : Jumlah Properti yang Sudah Laku vs. Belum Laku	76
Gambar IV.21 : Persentase Properti yang Sudah Laku per Kecamatan.....	77
Gambar IV.22 : Matriks Korelasi Fitur Numerik	78
Gambar IV.23 : Matriks Korelasi Seluruh Fitur	79
Gambar IV.24 : Variabel yang Digunakan Untuk Pelatihan Model.....	80
Gambar IV.25 : Preprocessor	81
Gambar IV.26 : Kode Membangun Model	83
Gambar IV.27 : Kode Inisiasi Kombinasi Hyperparameter Model pada Grid.	83
Gambar IV.28 : Kode Pelatihan Model	84
Gambar IV.29 : Perbandingan Evaluasi Model.....	86
Gambar IV.30 : RMSE Model pada Data Test.....	87
Gambar IV.31 : MAE Model pada Data Test.....	88
Gambar IV. 322 : R2 Model pada Data Test.....	88
Gambar IV.33 : Predicted vs Actual Price Random Forest.....	89
Gambar IV.34 : Residual Model Random Forest	89
Gambar IV.35 : Predicted vs Actual Price Regresi Linear	90
Gambar IV.36 : Residual Model Regresi Linear	90
Gambar IV.37 : Kode Untuk Mengambil Feature Importance Random Forest	91
Gambar IV.38 : Kode Untuk Mengambil Koefisien Regresi	94
Gambar IV.39 : Prediksi pada Sampel Data Test dengan Kedua Model	98
Gambar IV.40 : Prototype Web untuk Inference Model.....	100
Gambar IV.41 : Tab Tentang Model.....	100

DAFTAR PUSTAKA

- Ali, Z., & Bhaskar, Sb. (2016). Basic statistical tools in research and data analysis. *Indian Journal of Anaesthesia*, 60(9), 662. <https://doi.org/10.4103/0019-5049.190623>
- Alpaydm, E. (2021). *Machine Learning*. The MIT Press.
<https://doi.org/10.7551/mitpress/13811.001.0001>
- Bisong, E. (2019). Introduction to Scikit-learn. Dalam *Building Machine Learning and Deep Learning Models on Google Cloud Platform* (hlm. 215–229). Apress. https://doi.org/10.1007/978-1-4842-4470-8_18
- Boegershausen, J., Datta, H., Borah, A., & Stephen, A. T. (2022). Fields of Gold: Scraping Web Data for Marketing Insights. *Journal of Marketing*, 86(5), 1–20. <https://doi.org/10.1177/00222429221100750>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
<https://doi.org/10.1023/A:1010933404324>
- Cheng, V., Li, C.-H., Kwok, J. T., & Li, C.-K. (2004). Dissimilarity learning for nominal data. *Pattern Recognition*, 37(7), 1471–1477.
<https://doi.org/10.1016/j.patcog.2003.12.015>
- Diksha Bhola, Rachit Yadav, & Vimmi Malhotra. (2024). Introduction to Statistical Data Analysis. *International Journal of Advanced Research in Science, Communication and Technology*, 185–194.
<https://doi.org/10.48175/IJARSCT-17830>
- Galkanov, A. G. (2023). ABOUT INNOVATIVE METHODS OF NUMERICAL DATA AVERAGING. *RSUH/RGGU Bulletin. Series Information Science. Information Security. Mathematics*, 2, 81–101.
<https://doi.org/10.28995/2686-679X-2023-2-81-101>
- Gibson, D., Kleinberg, J., & Raghavan, P. (2000). Clustering categorical data: an approach based on dynamical systems. *The VLDB Journal The International Journal on Very Large Data Bases*, 8(3–4), 222–236.
<https://doi.org/10.1007/s007780050005>
- Grohe, M., Lucien Kaminski, B., Katoen, J.-P., & Lindner, P. (2021). Probabilistic Data with Continuous Distributions. *ACM SIGMOD Record*,

- 50(1), 69–76. <https://doi.org/10.1145/3471485.3471502>
- Gupta, P., & V., N. (2020). The Impact and Importance of Statistics in Data Science. *International Journal of Computer Applications*, 176(24), 10–14. <https://doi.org/10.5120/ijca2020920215>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York. <https://doi.org/10.1007/978-0-387-84858-7>
- Juliarto, R. (2021a, Maret 10). *Apa itu Activity Diagram? Beserta Pengertian, Tujuan, Komponen*. <https://www.dicoding.com/blog/apa-itu-activity-diagram/>
- Juliarto, R. (2021b, Mei 19). *Contoh Use Case Diagram Lengkap dengan Penjelasannya*. <https://www.dicoding.com/blog/contoh-use-case-diagram/>
- Khder, M. (2021). Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. *International Journal of Advances in Soft Computing and its Applications*, 13(3), 145–168. <https://doi.org/10.15849/IJASCA.211128.11>
- Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., & Team, J. D. (2016). *Jupyter Notebooks - a publishing format for reproducible computational workflows*. 87–90. <https://doi.org/10.3233/978-1-61499-649-1-87>
- Liddell, T. M., & Kruschke, J. K. (2015). Analyzing Ordinal Data: Support for a Bayesian Approach. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2692323>
- Mu'tashim, M. L., Muhayat, T., Damayanti, S. A., Zaki, H. N., & Wirawan, R. (2021). Analisis Prediksi Harga Rumah Sesuai Spesifikasi Menggunakan Multiple Linear Regression. *Informatik : Jurnal Ilmu Komputer*, 17(3), 238. <https://doi.org/10.52958/iftk.v17i3.3635>
- Oliphant, T. E. (2007). Python for Scientific Computing. *Computing in Science & Engineering*, 9(3), 10–20. <https://doi.org/10.1109/MCSE.2007.58>
- Omg. (2009). *An OMG ® Unified Modeling Language ® Publication OMG ® Unified Modeling Language ® (OMG UML ®) OMG Document Number: Date*. <https://www.omg.org/spec/UML/20161101/PrimitiveTypes.xmi>

- Padilah, T. N., & Adam, R. I. (2019). ANALISIS REGRESI LINIER BERGANDA DALAM ESTIMASI PRODUKTIVITAS TANAMAN PADI DI KABUPATEN KARAWANG. *FIBONACCI: Jurnal Pendidikan Matematika dan Matematika*, 5(2), 117. <https://doi.org/10.24853/fbc.5.2.117-128>
- Poslavskaya, E., & Korolev, A. (2023). Encoding categorical data: Is there yet anything “hotter” than one-hot encoding? *ArXiv*, *abs/2312.16930*. <https://doi.org/10.48550/arXiv.2312.16930>
- Powers, D., & Ailab. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *J. Mach. Learn. Technol*, 2, 2229–3981. <https://doi.org/10.9735/2229-3981>
- Qolbi, M. B. S., Puteh, T. N., Rivandi, R., & Rozikin, C. (2025). Prediksi Harga Rumah Di Jakarta Pusat Menggunakan Algoritma Machine Learning. *Jurnal Ilmu Komputer dan Bisnis*, 16(1), 16–24. <https://doi.org/10.47927/jikb.v16i1.840>
- Rafif Nauval Tuah Siregar, Vijay Sitorus, & Willy Pramudia Ananta. (2023). Analisis Prediksi Harga Rumah di Bandung Menggunakan Regresi Linear Berganda. *Journal of Creative Student Research*, 1(6), 395–404. <https://doi.org/10.55606/jcsrpolitama.v1i6.3038>
- Saurkar, A., Pathare, K., & Gode, S. (2018). *An Overview On Web Scraping Techniques And Tools*. <https://consensus.app/papers/an-overview-on-web-scraping-techniques-and-tools-saurkar-pathare/60d5109df7b85b5886cfdfa5ee9b4bdc/>
- Simpson, D. G. (1991). The Statistical Analysis of Discrete Data. *Technometrics*, 33(2), 242–242. <https://doi.org/10.1080/00401706.1991.10484818>