|  |
| --- |
| © Jurnal Nasional Teknik Elektro dan Teknologi Informasi  Karya ini berada di bawah Lisensi Creative Commons Atribusi-BerbagiSerupa 4.0 Internasional  DOI: 10.22146/jnteti.Nomor DOI |

House Prices Advanced Regression Techniques using XGboost

12S21009 Mikhael Janugrah Pakpahan1, 12S21010 Bobby Willy Siagian 2, 12S21011 Aldi Jeremy Simamora3

[Diserahkan: 12 Desember 2024, Direvisi: xx Januari 2024, Diterima: 1 Februari 2024]

(email: [Mikhaelpakpahan502@gmail.com](mailto:Mikhaelpakpahan502@gmail.com), [siagianbobby29@gmail.com](mailto:siagianbobby29@gmail.com), [aldyjeremysimamora@gmail.com](file:///C:\Users\Asus\Downloads\aldyjeremysimamora@gmail.com) )

**I**NTISARI ***—*** Dokumen ini merupakan format panduan bagi penulis untuk menulis naskah yang siap dipublikasikan di Jurnal Nasional Teknik Elektro dan Teknologi Informasi. Penulis harus mengikuti petunjuk yang diberikan dalam panduan ini. Penulis dapat menggunakan dokumen ini baik sebagai petunjuk penulisan maupun sebagai templat yang dapat langsung penulis isi dengan bagian-bagian naskah yang sesuai. Setiap kata di judul naskah harus diawali dengan huruf besar, kecuali pada kata-kata tertentu seperti kata penghubung. Hindari menulis rumus panjang dengan *subskrip* di judul, sedangkan rumus pendek yang mengidentifikasi elemen dapat dituliskan (misalnya, “Nd-Fe-B”). Nama lengkap penulis wajib dicantumkan di bawah judul. Nomor superskripdi belakang nama penulis menunjukkan afiliasi penulis. Intisari harus merupakan cerminan singkat tetapi komprehensif dari isi naskah. Penulisan persamaan, tabel, kutipan, dan referensi dalam Intisari harus dihindari. Tuliskan Intisari dalam bahasa Indonesia, sepanjang 200-250 kata. Intisari berisi paling tidak empat bagian. Bagian pertama adalah pengantar yang meliputi seperempat bagian dari intisari, kemudian diikuti tujuan penulisan naskah. Bagian ketiga adalah metodologi singkat dan bagian terakhir merupakan penjelasan tentang hasil penelitian secara keseluruhan.

KATA KUNCI — Tuliskan 4-8 kata kunci Anda di sini. Kata kunci dipisahkan dengan koma dan masing-masing kata diawali dengan huruf kapital.

PENDAHULUAN

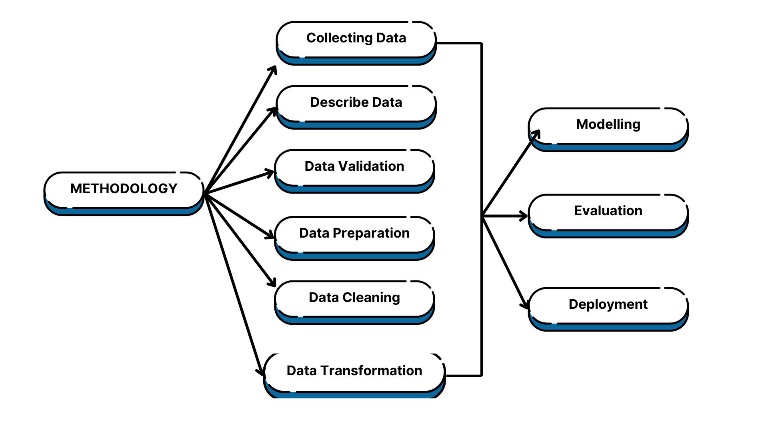
Kemajuan teknologi dan melimpahnya data telah membuka peluang besar untuk memecahkan berbagai masalah kompleks, termasuk dalam menentukan harga rumah. Prediksi harga rumah sangat penting bagi berbagai pihak, seperti pengembang properti, investor, dan pembeli, karena dapat memberikan gambaran yang akurat mengenai nilai properti berdasarkan berbagai faktor yang memengaruhinya. Salah satu metode yang saat ini banyak digunakan dalam bidang ini adalah XGBoost (Extreme Gradient Boosting), algoritma pembelajaran mesin berbasis ensemble yang dikenal memiliki performa tinggi dalam menangani data kompleks.

Studi ini bertujuan membangun model prediksi harga rumah yang andal menggunakan XGBoost. Prosesnya mencakup eksplorasi data, pemrosesan, pemilihan fitur, pelatihan model, dan evaluasi kinerja menggunakan metrik seperti Root Mean Square Error (RMSE). Dengan memanfaatkan pendekatan ini, diharapkan model yang dihasilkan tidak hanya akurat, tetapi juga mampu memberikan wawasan berharga bagi pengambilan keputusan yang lebih baik dalam industri properti.

METODOLOGI

Bab ini menyajikan metode yang digunakan dalam penelitian untuk mengembangkan model prediksi harga rumah menggunakan algoritma XGBoost. Penyusunan metodologi ini bertujuan untuk memastikan pendekatan yang sistematis dan terstruktur, sehingga hasil yang diperoleh dapat diandalkan dan memiliki validitas ilmiah.

Proses pengumpulan data dimulai dengan menggunakan dataset publik yang relevan, seperti Kaggle House Prices dataset, yang mencakup berbagai atribut penting. Data yang diperoleh kemudian diproses dengan hati-hati, termasuk langkah-langkah pembersihan, penanganan data yang hilang, dan normalisasi, guna memastikan kualitas data yang digunakan untuk pelatihan model.



Gambar 1. Metodologi

* ****Collecting Data****

Data untuk studi ini diambil dari dataset publik yang tersedia, seperti Kaggle House Prices dataset. Dataset ini mencakup berbagai fitur seperti ukuran rumah, lokasi, tahun pembangunan, dan kondisi rumah.

* ****Describe Data****

Langkah ini mencakup eksplorasi data untuk memahami struktur dan karakteristik dataset. Tahapan utama meliputi:

1. Pembersihan Data (Data Cleaning): Mengatasi data yang hilang, duplikasi, atau inkonsistensi.
2. Normalisasi: Menstandarkan data sehingga berada dalam skala yang seragam untuk mencegah dominasi fitur tertentu.
3. Pemilihan Fitur: Fitur yang relevan dipilih berdasarkan analisis statistik untuk memastikan bahwa model hanya menggunakan variabel yang signifikan, sehingga mengurangi risiko overfitting.

* ****Data Validation****

Pada tahap ini, algoritma pembelajaran mesin XGBoost digunakan untuk melatih model. Proses ini mencakup:

1. Optimasi Parameter: Menggunakan metode Grid Search untuk menemukan kombinasi parameter terbaik yang meningkatkan performa model.
2. Validasi Data: Model dilatih dan diuji untuk memastikan bahwa prediksi yang dihasilkan dapat diterapkan pada data baru dengan akurasi tinggi..

* ****Data Preparation****

Persiapan data melibatkan langkah-langkah untuk memastikan data siap digunakan dalam model. Tahapan ini mencakup:

1. Evaluasi Model: Mengukur performa model dengan membandingkan prediksi dengan data aktual menggunakan metrik seperti RMSE (Root Mean Squared Error).
2. Validasi Silang: Membagi dataset menjadi beberapa bagian untuk

memastikan konsistensi performa model di berbagai subset data.

* Data Cleaning

Fokus pada membersihkan data dari elemen-elemen yang tidak relevan atau bermasalah, seperti:

1. Mengisi nilai-nilai yang hilang.
2. Menghapus duplikasi.
3. Memperbaiki data yang tidak valid atau salah format.

* Data Transformation

Transformasi data dilakukan untuk mempermudah analisis dan pelatihan model. Contohnya:

1. Mengubah data kategorikal menjadi format numerik (misalnya dengan one-hot encoding).
2. Menggunakan teknik scaling atau normalization untuk menyamakan rentang nilai.

* Modeliing

Pada tahap ini, model prediktif dibuat menggunakan algoritma XGBoost. Tahapan penting meliputi:

1. Membagi dataset menjadi data latih dan data uji.
2. Melatih model dengan data latih.
3. Mengoptimalkan performa model melalui tuning parameter.

* Evaluation

Model yang dibangun dievaluasi untuk menilai performanya menggunakan metrik seperti:

1. RMSE: Mengukur kesalahan rata-rata dalam prediksi model.
2. Akurasi Validasi Silang: Memastikan bahwa model bekerja konsisten pada data yang berbeda.  
   Hasil evaluasi digunakan untuk memahami kekuatan dan kelemahan model.

* Deployment

Model yang telah divalidasi dan dievaluasi kemudian diterapkan ke dalam lingkungan produksi. Tahapan ini mencakup:

1. Integrasi model ke dalam aplikasi atau sistem yang akan menggunakannya.
2. Pemantauan performa model secara berkala untuk memastikan hasil tetap akurat.
3. Pembaruan model jika diperlukan berdasarkan data baru.

HASIL DAN ANALISIS

Pada bab ini, hasil dari pengolahan data, pelatihan model, dan evaluasi model akan dibahas secara rinci. Proses ini mencakup eksplorasi data, pemilihan fitur, pelatihan model XGBoost, dan evaluasi performa model dalam memprediksi harga rumah. Analisis hasil dilakukan untuk memahami seberapa baik model dapat memprediksi nilai properti berdasarkan fitur yang tersedia dalam dataset.

A. Eksplorasi Data

Eksplorasi data dilakukan untuk memahami karakteristik dataset, termasuk distribusi nilai variabel, jumlah data yang hilang, dan hubungan antar variabel. Dataset terdiri dari tiga file utama:

1. **train.csv** - berisi data historis properti dengan fitur dan harga rumah (SalePrice) sebagai variabel target.
2. **test.csv** - berisi data properti tanpa informasi harga rumah untuk diuji.
3. **sample\_submission.csv** - template untuk mengunggah hasil prediksi harga rumah.
4. LABEL DAN JUDUL TABEL

Tabel diberi label menggunakan angka Romawi huruf kapital. Label tabel ditulis rata tengah (*centered*) dan dalam Times New Roman biasa berukuran 8. Setiap awal kata dalam judul tabel menggunakan huruf kapital, kecuali untuk kata hubung, preposisi, partikel, dan sejenisnya, seperti yang tercantum pada bagian III-B. Label dan judul tabel ditulis menggunakan Small Caps. Label dan judul tabel ditempatkan sebelum tabel terkait, seperti yang ditunjukkan pada Tabel I.

1. DESKRIPSI STATISTIK VARIABEL

Tabel berikut menunjukkan deskripsi statistik dari beberapa variabel penting dalam dataset:

Tabel 1. Statistik Variabel

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variabel | Mean | Median | Min | Max | Std. Dev |
| LotArea | 10517.0 | 9478.5 | 1300 | 215245 | 9981.3 |
| YearBuilt | 1971.3 | 1973.0 | 1872 | 2010 | 30.2 |
| OverallQual | 6.1 | 6.0 | 1 | 10 | 1.4 |
| OverallCond | 5.6 | 5.0 | 1 | 10 | 1.1 |
| SalePrice | 180921.2 | 163000.0 | 34900 | 7550000 | 79442.1 |

Data menunjukkan distribusi yang beragam untuk setiap variabel, dengan beberapa outlier terutama pada variabel LotArea dan SalePrice.

1. PENANGANAN DATA HILANG

Beberapa variabel dalam dataset memiliki nilai yang hilang, seperti LotFrontage. Nilai yang hilang ditangani dengan teknik imputasi, seperti:

* **LotFrontage**: Menggunakan median berdasarkan Neighborhood.
* **MasVnrArea** dan **GarageYrBlt**: Mengisi dengan 0 untuk properti tanpa fitur ini.

D. PEMILIHAN FITUR

Fitur yang digunakan dalam pelatihan model dipilih berdasarkan relevansi terhadap harga rumah dan analisis korelasi. Beberapa fitur utama yang digunakan:

* **LotArea**: Ukuran tanah
* **Neighborhood**: Lokasi properti
* **YearBuilt**: Tahun konstruksi selesai
* **OverallQual**: Kualitas keseluruhan rumah
* **GrLivArea**: Luas ruang lantai di atas tanah

Korelasi antara fitur dan variabel target (SalePrice) dianalisis menggunakan heatmap korelasi. Fitur dengan korelasi tinggi seperti OverallQual (0.79) dan GrLivArea (0.70) diprioritaskan.

E. PELATIHAN MODEL

Model XGBoost digunakan untuk memprediksi harga rumah. Dataset dibagi menjadi data pelatihan (train) dan data validasi untuk menghindari overfitting. Parameter utama yang digunakan dalam model meliputi:

* **n\_estimators**: 100
* **learning\_rate**: 0.1
* **max\_depth**: 6

EVALUASI MODEL

Model dievaluasi menggunakan metrik berikut:

* Mean Absolute Error (MAE)
* Root Mean Squared Error (RMSE)
* R-squared (R²)

Hasil evaluasi pada data validasi:

|  |  |
| --- | --- |
| **Metrik** | **Nilai** |
| MAE | 17035.2 |
| RMSE | 24321.5 |
| R2 | 0.87 |

E. PELATIHAN MODEL

F. ANALISIS HASIL

KESIMPULAN

Penggunaan serial koma (koma Oxford) sangat dianjurkan: “A, B, dan C,” bukan “A, B dan C.” Kata-kata bentuk terikat, seperti pre-, multi-, dan sub-, ditulis bersambung dengan kata yang mengikutinya. Misalnya, “multifungsi.”

REFERENSI

M. Metev dan P. Veiko, Laser Assisted Microtechnology, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998, doi: 10.1007/978-3-642-87271-6.

1. J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, Vol. 61, doi: 10.1007/978-1-4612-3688-7.
2. S. Zhang, C. Zhu, J.K.O. Sin, dan P.K.T. Mok, “A Novel Ultrathin Elevated Channel Low-temperature Poly-Si TFT,” *IEEE Electron Device Lett.*, Vol. 20, hal. 569–571, Nov. 1999, doi: 10.1109/55.798046.
3. M. Wegmuller, J.P. von der Weid, P. Oberson, dan N. Gisin, “Highresolution Fiber Distributed Measurements with Coherent OFDR,” *Proc. ECOC’00*, 2000, paper 11.3.4, hal. 109.
4. R.E. Sorace, V.S. Reinhardt, dan S.A. Vaughn, “High-speed Digital-to-RF Converter,” U.S. Patent 5 668 842, 16 Sep. 1997.
5. (2002) The IEEE website. [Online], http://www.ieee.org/, tanggal akses: 16-Sep-2014.
6. M. Shell. (2002) IEEEtran homepage on CTAN. [Online], http://www.ctan.org/tex-a/rchive/macros/latex/contrib/supported/IEEEtran/, tanggal akses: 16-Sep-2014.
7. *FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.
8. “PDCA12-70 data sheet,” Opto Speed SA, Mezzovico, Switzerland.
9. A. Karnik, “Performance of TCP Congestion Control with Rate Feedback: TCP/ABR and Rate Adaptive TCP/IP,” M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.
10. J. Padhye, V. Firoiu, dan D. Towsley, “A Stochastic Model of TCP Renocongestion Avoidance and Control,” Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. hal. 99-02, 1999.
11. *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 1997.

LAMPIRAN (OPSIONAL)

Lampiran, jika diperlukan, diletakkan setelah referensi. Lampiran berisi materi tambahan yang bukan merupakan bagian penting dari naskah tetapi dapat membantu dalam memberikan pemahaman yang lebih komprehensif tentang masalah penelitian dan/atau informasi yang terlalu rumit untuk dimasukkan ke dalam badan naskah. Poin kunci yang perlu diingat ketika menulis lampiran adalah bahwa informasi tersebut tidak penting; jika dihapus, naskah itu masih bisa dimengerti.

Lampiran ditambahkan ketika penambahan informasi dalam badan naskah akan menyebabkan naskah menjadi tidak terstruktur dengan baik atau terlalu panjang dan rinci. Selain itu, lampiran digunakan untuk memastikan dimasukkannya materi yang bermanfaat, mendukung, atau penting; jika tidak, akan mengacaukan atau memecah alur naskah, atau akan mengganggu pembaca. Lampiran juga dapat digunakan untuk meletakkan gambar atau tabel yang terlalu besar dan memenuhi satu halaman naskah.