eISSN : 2621-4970
pISSN : 2621-4962

Volume 8 | No 2
Juli | **2025**

Penerapan Algoritma Random Forest Untuk Prediksi Biaya Kontruksi Berbasis Web

Dui Puspitasari[1], Noora Qotrun Nada[2], Aris Tri Jaka Harjanta[3]
Fakultas Teknik dan Informatika, Program Studi Informatika
Universitas PGRI Semaranag
duipuspitasari123@gmail.com

Abstrak

Estimasi biaya proyek konstruksi sangat penting untuk menjamin efektivitas dan ketepatan perencanaan anggaran, kompleksitas proyek dan banyaknya variabel yang terlibat seringkali menyulitkan perusahaan konstruksi dalam menghasilkan estimasi biaya yang akurat. Tujuan dari penelitian ini adalah menggunakan data historis dan algoritma *random forest regression* untuk memperkirakan biaya proyek bangunan. Karena kapasitasnya untuk mengelola data yang rumit, meminimalkan *overfitting*, dan meningkatkan akurasi prediksi, pendekatan *random forest* dipilih. Model Random Forest digunakan untuk mengumpulkan, membersihkan, dan melatih data proyek sebelum dimasukkan ke dalam sistem informasi daring. Hasil pengujian menunjukkan tingkat akurasi model yang tinggi dalam esti masi biaya, tim proyek dan manajemen dapat mengakses data estimasi dengan cepat dan efektif berkat teknologi ini. Secara keseluruhan, penggunaan algoritma random forest dalam sistem berbasis web menawarkan cara yang fleksibel dan tepat untuk mendukung proses estimasi biaya proyek konstruksi.

Kata kunci: Konstruksi, Machine Learning, Prediksi, Proyek, Random Forest.

Abstract

Construction project cost estimation is crucial to ensure the effectiveness and accuracy of budget planning. Project complexity and the numerous variables involved often make it difficult for construction companies to produce accurate cost estimates. The purpose of this study is to use historical data and the random forest regression algorithm to estimate the cost of a building project. Due to its capacity to handle complex data, minimize overfitting, and improve prediction accuracy, the random forest approach was chosen. The random forest model was used to collect, clean, and train project data before being input into an online information system. Test results demonstrated a high level of model accuracy in cost estimation, and project teams and management were able to access the estimated data quickly and effectively thanks to this technology. Overall, the use of the random forest algorithm in a web-based system offers a flexible and appropriate way to support the cost estimation process of construction projects..

Keywords: *Construction, Machine Learning, Prediction, Project, Random Forest.*

PENDAHULUAN

Pengelolaan biaya proyek merupakan salah satu aspek krusial dalam industri konstruksi. Biaya proyek mencakup berbagai komponen penting seperti tahapan pelaksanaan, desain, analisis biaya material, tenaga kerja, hingga logistik. Tantangan dalam mengelola semua komponen ini secara efisien seringkali dihadapi oleh perusahaan konstruksi, khususnya dalam proyek-proyek berskala

besar seperti infrastruktur transportasi. Namun, ketika jumlah dan kompleksitas proyek meningkat, perusahaan konstruksi kerap menghadapi tantangan dalam mengelola anggaran proyek secara efisien dan akurat. Penggunaan estimasi biaya yang tidak akurat dan tanpa baseline yang valid dapat menyebabkan pembengkakan anggaran, pemborosan, serta menurunnya kredibilitas system estimasi[1]. Oleh karena itu, aplikasi yang dapat membantu menghitung estimasi biaya proyek konstruksi adalah dibutuhkan agar perusahaan dapat mengetahui biaya yang tepat dan meminimalkan kelebihan biaya risiko[2].

Terkait estimasi biaya proyek konstruksi, berbagai pendekatan tradisional, termasuk estimasi berdasarkan pengalaman, perhitungan unit price, dan model analisis linier regresi dibanjir belum berganti[3]. Meskipun begitu, cara ini tidak dapat menanggung beban pada data dalam proporsi besar, terutama ketika sejumlah besar variabel saling berkaitan atau diatur secara non-linier. Dengan kata lain, pendekatan baru dan adaptif cenderung dikelola.

Machine learning ialah cabang dari kecerdasan buatan yang memungkinkan sistem belajar dari pengalaman, yaitu data machine learning digunakan ketika terlalu rumit atau sulit bagi manusia untuk menulis algoritma dasar. Salah satu metode machine learning yang efektif untuk masalah regresi adalah *Random Forest Regression.* Random Forest pertama kali diperkenalkan oleh Breiman (2001), sebagai penyempurnaan dari metode decision tree dengan memanfaatkan konsep bagging *(bootstrap aggregating)*[4]. Random Forest memiliki keunggulan dalam mengurangi *overfitting*, meningkatkan akurasi prediksi, serta mampu menangani data yang mengandung missing value maupun outlier[5].

Algoritma ensemble yang disebut Random Forest didasarkan pada pohon keputusan yang mengoptimalkan akurasi dan stabilitas dengan cara membentuk sejumlah pohon keputusan yang berasal dari berbagai pohon keputusan dengan seleksi subset data yang diambil secara acak yang dikelompokkan kembali dengan cara menggabungkan pohon-pohon tersebut sebagai estimasi prediksi terakhirnya[4]. Dalam setiap pohon pembentukan, juga menggunakan variabel secara acak sebagai titik split atau penentu dan digunakan beberapa tokoh dengan korelasi yang low sehinga membuat Random Forest lebih general[6]. Nantinya, Random Forest dapat menghasilkan general model yang optimal dengan mengatasi data yang kompleks dan *fitesh* korelasi ataupun autocorrelation yang tinggi pada tiap variabel konvensional[7].

Penelitian ini bertujuan untuk mengeksplorasi penerapan algoritma Random Forest dalam memprediksi biaya proyek konstruksi, di mana kasus yang dimasukan adalah PT Langgeng Dadi Pratama. Model yang akan diperoleh menggunakan data historis PT Langgeng Dadi Pratama dibenarkan agar perusahaan tersebut memperoleh bahasa untuk informasi yang sangat dibutuhkan bagi penyusunan anggaran dan pengembangan harga atau biaya. Dengan web yang dikembangkan, seluruh pihak yang terlibat dalam lingkaran transaksional proyek dan biaya memiliki akses dan kemampuan yang setara dalam pengelolaan informasi, diharapkan PT Langgeng Dadi Pratama dapat meramalkan data biaya proyek masa

lalu dan pengalaman yang ingin jika telah digunakan.

Rumusan Masalah

Bagaimana penerapan algoritma Random Forest dapat digunakan untuk memprediksi estimasi biaya proyek konstruksi secara akurat melalui sistem berbasis web?

Tinjauan Pustaka

Prediksi biaya proyek konstruksi sangat penting karena jika dilakukan secara tidak tepat dapat mengakibatkan pemborosan anggaran dan keterlambatan proyek. Berbagai penelitian telah dilakukan untuk menghasilkan prediksi biaya yang lebih akurat, seperti menggunakan algoritma machine learning. Farhanuddin [8] membandingkan model *random forest regression* dan *multiple linear regression* untuk memperkirakan biaya proyek sistem informasi. Dengan menggunakan model *random forest regression*, mereka memperoleh akurasi sebesar 81,6%, yang lebih baik daripada metode *multiple linear regression*. Namun, keduanya masih memiliki kesamaan di sistem informasi, tetapi beberapa aspek terutama berhubungan dengan bagaimana atau tugas apa yang dapat dilakukan oleh metode itu sendiri, yang dipertimbangkan untuk studi konstruksi fisik atau proyek berbasis web ini.

Menggunakan data dari web scraping, Lestari dan Astuti [9] menggunakan algoritma *random forest* untuk memperkirakan nilai rumah. Mereka berhasil membuat sistem prediksi berbasis web dengan akurasi 85,29%. Metode ini menunjukkan penggunaan algoritma *random forest* memiliki potensi untuk diimplementasikan dalam estimasi berbasis web, termasuk untuk biaya konstruksi, maupun pada bidang lainnya. Amelia dan Kurniawan [10] menggunakan Random Forest untuk menentukan kebutuhan produk dan penjualan di beku toko. Akurasi prediksi mereka mencapai 83%. Studi ini menunjukkan efektivitas Random Forest dalam menangani data kompleks dan memungkinkan implementasi berbasis web untuk penelitian bisnis. Selain itu, Yuliana dan Yuni [1] menyoroti pentingnya manajemen risiko dalam proses estimasi biaya konstruksi. Mereka mengidentifikasi 35 risiko potensial yang dapat memengaruhi akurasi estimasi, seperti kesulitan membaca dokumen tender, volume pekerjaan, dan keterlambatan pembayaran. Namun, penelitian ini tidak menggunakan analisis prediktif berdasarkan data historis seperti *random forest*.

Random Forest adalah algoritma proses pembelajaran ensemble terdiri dari beberapa pohon keputusan, yang dilatih pada tugas yang sama atau berbeda, dan teknik optimasi atau komposit, yang sama atau berbeda. Ketika algoritma diterapkan dalam konteks regresi, prediksi akhir adalah rata-rata pesanan dari pohon yang dihasilkan. Keuntungan *random forest* membaca tidak memuncak, menahan untuk bekerja pada data dengan banyak fitur, dan dapat dioperasikan pada data yang proporasi tidak linier[8].

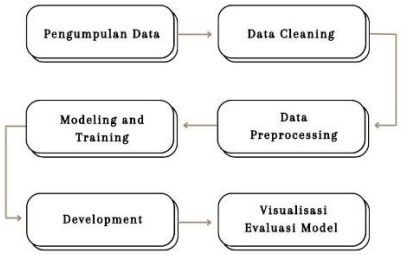
Estimasi biaya sendiri dapat didefinisikan sebagai prosedur yang mengidentifikasi dan memperkirakan biaya total yang dibutuhkan untuk melalui pekerjaan anggaran proyek. Estimasi biaya di tingkat perencanaan sangat luar biasa rentan terhadap beberapa jenis risiko

utama, karena proses ini sering berjalan sangat cepat di atas minggu yang terbatasi oleh pembatasan sumber daya. Sebagian besar estimator dilibatkan dalam jumlah pekerjaan ini yang mengalami kesulitan melaksanakan tugasnya tanpa kesalahan. Karena semakin sulit untuk membuat sistem pelaporan estimasi biaya yang benar, penyelesainya terletak pada data historis dan pemodelan algoritme prediktif. Sistem berbasis web untuk memprediksi biaya dapat digunakan hampir oleh pengguna untuk membuat sebuah estimasi biaya dalam mode interaktif, kolaboratif, dan komitmen data aktual atau disebut sebagai real-time data. Sistem berbasis web ditambah algoritma prediktif mungkin membuat operasional dan efisiensi keputusan strategi ditingkatkan[9].

Beberapa penelitian telah dirangkum kita dapat menyimpulkan bahwa algoritma random forest kita digunakan untuk tujuan pekerjaan prediksi kita dalam beberapa pekerjaan. Namun, penggunaan dalam bentuk pembuatan data primer berupa sistem prediksi biaya berbasis web ini masih relatif terbatas. Untuk menutup celah itu, maka penelitian ini akan dilakukan.

METODE PENELITIAN

Dalam penelitian ini, objek yang digunakan adalah data historis proyek konstruksi yang mencakup berbagai atribut penting, seperti jenis bangunan, kategori pekerjaan, lokasi, durasi proyek, jumlah tenaga kerja, serta luas bangunan. Data ini diperoleh dari database internal perusahaan konstruksi dan diolah untuk menghasilkan prediksi biaya konstruksi. Adapun penelitian ini terbagi menjadi beberapa tahapan, yang akan dapat dilihat pada gambar 1 di bawah ini.



Gambar 1. Tahapan Penelitian

Pengumpulan Data

Pengumpulan data merupakan tahap awal dari penelitian ini yang akan digunakan sebagai landasan utama pengembangan sistem prediksi biaya konstruksi selanjutnya. Data dikumpulkan langsung dari PT Langgeng Dadi Pratama yang merupakan perusahaan konstruksi sekaligus mitra penelitian. Karena sumber data berasal langsung dari lembaga terkait dan tidak dipublikasikan atau diperoleh dari pihak ketiga, maka data tersebut dianggap sebagai data primer.

Data Cleaning

Setelah seluruh data proyek terkumpul, dilakukan pembersihan data, kolom nama dihapus karena hanya berisi deskipsi proyek yang tidak dibutuhkan dalam prediksi. Model hanya menggunakan kolom-kolom penting misal jenis bangunan, kategori, lokasi, durasi proyek, jumlah tenaga kerja, luas bangunan, dan nilai kontrak. Seluruh data dipastikan tidak mengandung nilai kosong karena seluruh kolom terisi, maka tidak dilakukan metode pengisian atau penghapusan baris. Data dengan nilai tidak logis juga dihapus, misal nilai kontrak atau luas bangunan yang bernilai nol, untuk memastikan tidak memengaruhi hasil prediksi[11].

Nilai kontrak memiliki keragaman yang sangat besar karena memegang log1p digunakan, Ini membantu model untuk memahami pola data dengan lebih baik dan mengurangi dampak dari data yang berada di 'ekor' yang jauh. Para penulis menggunakan expm1 untuk mengembalikan log dari hasil prediksi y_log ke dalam bentuk aslinya. Akhirnya, mereka menyesuaikan jenis data mereka. Mereka mengonversi numerical data seperti durasi dan luas ke angka yang sebenarnya dan categorical termasuk type_bangunan, kategori, dan lokasi dikodekan dalam bentuk angka dengan menggunakan pendekatan *one-hot encoding.*

Data Preprocessing

Proses pra-pemrosesan data yang perlu dilakukan sebelum memproses data adalah memastikan data yang akan kita gunakan dalam pemodelan sudah bersih, konsisten dan siap untuk dianalisis[12]. Adapun proses dari pra-pemrosesan hasil pembersihan dilakukan, yang pertama melakukan pembersihan kolom yang melanggar aturan, misalnya kolom nama, modifikasi kolom dan digunakan ke dalam format yang benar. Berikutnya melakukan identifikasi fitur dalam dataset, memisahkan dataset dan fitur dibagi menjadi dua kelompok yaitu fitur kategoris adalah tipe bangunan, kategori, dan lokasi serta fitur numerisnya adalah durasi (dalam minggu), dan jumlah tukang. Fitur kategori mengonversi ke dalam representasi binari numerik untuk binaraisasi menggunakan metode *One-Hot Encoding*, dan fitur numeric dinormalisasi juga dengan metode *MinMaxScaler* dioperasikan agar fitur dengan skala yang lebih tinggi[13]. Untuk sistematis menerapkan operasi tersebut ke dalam kode kami menggunakan Pipeline dan *ColumnTransformer* dari library scikit-learn.hal ini membuat kami lebih mudah dalam melakukan operasi pra-pemrosesan kembali.

Modeling and Training

Setelah data siap digunakan, dataset tersebut akan dibahagi menjadi dua data set, data latih dan data uji. Pembagian data set dipisahkan dengan menggunakan fungsi train_test_split dari pustaka *scikit-learn* dengan proporsi 75% data latih dan 25% data uji. Selain itu, untuk kepentingan konsistensi hasil pembagian jika setiap pelatihan diulang, parameter *random_state=42* dijadikan nilai acuan tetap[14].Untuk menghindari potensi data leakage dan menjaga proses pelatihan agar berjalan secara efisien dan terstandarisasi, dibangun *pipeline* menggunakan modul *Pipeline* dari *scikit-learn*. Pipeline yang dibuat menggabungkan tahapan pemrosesan data antara satu dan lainnya secara terintegrasi, mulai dari pra-pemrosesan hingga feature engineering dan pembuatan prediksi model. Proses pra-pemrosesan antara lain transformasi fitur kategorikal dan numerik dengan jenis transformasi yang berbeda, seperti encoding dan scaling. Pengujian dilakukan menggunakan algoritma *Random Forest Regressor*. Model Random Forest dikonfigurasi dengan estimator sebanyak 100 pohon keputusan, *random_state*=42 untuk keseragaman hasil, dan n_jobs=-1 agar proses komputasi berjalan paralel[15]. Setelah dirancang, proses pelatihan model dijalankan dengan memanggil fungsi fit pada posisi data latih. pendekatan ini memungkinkan seluruh transformasi data berjalan secara otomatis dan konsisten setiap kali model dijalankan.

Development

Tahap pada jenis ini adalah pengembangan sistem prediksi biaya konstruksi yang diimplementasikan dalam bentuk aplikasi web. Sistem diterjemahkan menggunakan bahasa Python dengan bantuan library Flask sebagai backend, HTML dan Bootstrap sebagai tampilan antarmuka[16]. Model prediksi dibuat dengan menggunakan algoritma Random Forest dan tersimpan dalam format.joblib, model ini sudah diintegrasikan dengan pengguna dengan model di atas.

Pengguna hanya perlu mengisi formulir berdasarkan tipe bangunan, tempat yang dipilih, durasi waktu proyek, jumlah pekerja, dan ukuran gedung. Selanjutnya, data tersebut akan diproses dan hasil perhitungan biaya akan ditunjukkan dalam bentuk mata uang Rupiah. Selain itu, hasil perhitungan akan disimpan ke dalam database MySQL dan pengguna dapat meninjau sejarah perhitungan yang pernah dikerjakan. Dengan adanya sistem ini, pengguna dapat membantu proses estimasti biaya proyek menjadi lebih mudah, cepat dan efisien tanpa perhitungan manual yang harus diperhatungkan sendiri.

Visualisasi Evaluasi Model

Setelah dilatih, model dievaluasi untuk mengukur akurasi dalam memprediksi biaya proyek konstruksi. Beberapa metrik evaluasi yang digunakan adalah Mean Absolute Error dan Root Mean Squared Error yang mencari rata-rata selisih antara hasil prediksi dan nilai sebenarnya[17]. Semakin kecil MAE dan RMSE, semakin baik model dalam prediksi. Selain itu, R-squared digunakan untuk mengetahui proporsi variasi yang dapat dijelaskan oleh model. Paling penting, nilai R-squared yang tinggi menunjukkan model yang baik dalam prediksi. Metrik lain yang digunakan adalah *Mean Absolute Percentage Error* untuk memperkirakan eror prediksi dalam bentuk persentase yang mudah dipahami, terutama terkait keuangan[18].

Untuk membantu pemahaman, dilakukan juga visualisasi hasil prediksi. *Grafik Actual vs. Predicted* digunakan untuk membandingkan antara nilai asli dengan hasil prediksi[19]. Jika titik-titik dalam grafik berada dekat dengan garis lurus, artinya prediksi cukup akurat. Selanjutnya, *residual plot* digunakan untuk melihat selisih antara nilai prediksi dan nilai asli. Jika residual tersebar secara acak, maka model dianggap tidak memiliki bias. Terakhir, ditampilkan juga *feature importance*, yaitu grafik yang menunjukkan fitur-fitur yang paling berpengaruh terhadap hasil prediksi. Misalnya, fitur seperti luas bangunan atau durasi proyek cenderung memberikan pengaruh besar dalam menentukan nilai kontrak.

HASIL DAN PEMBAHASAN

Tujuan dari penelitian ini adalah untuk memahami sistem prediksi biaya konstruksi dengan algoritma *Random Forest Regressor* yang terintegrasi pada aplikasi berbasis web. Sistem ini untuk membantu pihak manajemen proyek dalam melakukan prediksi nilai kontrak berdasarkan parameter teknis proyek tersebut.

Pengumpulan Data

Dataset yang digunakan dari 39 data proyek konstruksi langsung dari PT Langgeng Dadi Pratama. Terdapat data proyek konstruksi asli dari beragam proyek yang tersebar di berbagai wilayah

Indonesia. Distribusi jenis bangunan menunjukkan proyek bangunan memiliki total sebanyak 25 entri, diikuti rumah sebesar 8 entri dan instalasi 6 entri. Dari kategorisasi jenis pekerjaan, proyek destruktif selama enam bulan memiliki 25 data terbanyak, diikuti renovasi empat bulan 10 data dan konstruksi dua bulan 4 data. Lokasi proyek terbanyak dari Jabodetabek 30 data, Jawa Tengah 5 data dan lainnya tersebar di Kalimantan, Sumatera dan Jawa Barat satu data. Untuk segi tahun, jumlah data proyek tersebar sepanjang tahun efisien dengan skala tahun 2020, 2021, 2022, 2023, dan 2024 merupakan proporsi yang seimbang.

Data yang bervariasi mengenai durasi, jumlah karyawan, serta nilai dan area proyek menunjukkan bahwa kumpulan data cukup bervariasi dalam hal skala dan jenis pekerjaan konstruksi. Ini merupakan aspek penting dalam pelatihan model pembelajaran mesin karena memungkinkan model mengenali pola dalam berbagai jenis proyek dengan lebih akurat dan menghindari bias dalam jenis pekerjaan atau area kumpulan data tertentu.

Data Cleaning

Setelah Model Random Forest dilatih menggunakan data yang dikumpulkan sebelumnya, dan sistem dievaluasi menggunakan data pengujian baru, hasil prediksi ditampilkan dalam bentuk rupiah dan dapat dilihat langsung melalui tampilan antarmuka web. Nilai prediksi diproses kembali ke situs web setelah dievaluasi oleh fungsi eksponensial eksponital expm1, kembali ke nilai sebenarnya. Beberapa metrik evaluasi digunakan untuk menguji model termasuk *Mean Absolute Error, Mean Squared Error, dan R Squared.* Berdasarkan kinerja model tersebut, hasil model Random Forest cukup bagus dengan nilai R^2 mendekati 1. Sejauh ini, ini menunjukkan bahwa sebagian besar variasi dari nilai kontrak nyata dapat dijelaskan oleh model. Grafik berikut juga menunjukkan visualisasi titik hasil prediksi model. Di grafik data latih sebelumnya, titik cenderung mengikuti garis ideal. Ini menunjukkan bahwa model dapat menganalisis data secara akurat.

Di grafik data pengujian, meskipun sebagian prosesi ekor memotong, sejauh ini, titik konvergen berada pada nilai aktual, yang menunjukkan bahwa model dapat diandalkan dan masih berkinerja baik dalam hal data yang belum pernah dikenal. Terlebih lagi, sistem menyimpan data prediksi ke dalam basis data sesuai dengan riwayatnya sehingga pengguna dapat melihatnya kembali.

Hasil ini berguna untuk pengguna, terutama manajer, untuk melacak dan menganalisis data historis. Secara keseluruhan, hasil yang diperoleh menunjukkan bahwa model *Random Forest* sudah sesuai untuk memperkirakan biaya secara otomatis dan cepat. Catatan bahwa hasil akhirnya hanya bersifat perkiraan, itu selalu harus diatur dan mungkin tidak untue keputusan proyek. Validasi lapangan dan pertimbangan teknis diperlukan dalam setiap keadaan kasus

Data Preprocessing

Proses pra-pemrosesan data berperan penting dalam meningkatkan mutu masukan (input) bagi model prediksi biaya konstruksi. Tahapan pra-pemrosesan data dimulai dengan pembersihan kolom dan validasi nilai target, kolom bernama "nama;" disesuaikan menjadi nama untuk

memastikan konsistensi atribut. Selain itu, data dengan nilai kontrak sebesar nol dihapus agar tidak menyebabkan error saat dilakukan transformasi logaritmik.

Selanjutnya, nilai target nilai_kontrak ditransformasikan menggunakan fungsi log1p() untuk mengurangi skewness dan memperkecil rentang data, sehingga memudahkan model dalam mengenali pola. Data fitur dibagi menjadi dua kategori: fitur numerik dan fitur kategorikal. Fitur numerik seperti durasi_minggu, jumlah_tukang, dan luas_meter_persegi dilakukan normalisasi menggunakan *Min-Max Scaler*. Sementara itu, fitur kategorikal seperti type_bangunan, kategori, dan lokasi dikonversi menjadi bentuk numerik menggunakan *One-Hot Encoding*. Seluruh proses pra-pemrosesan ini diintegrasikan dalam pipeline *ColumnTransformer*, sehingga dapat berjalan secara otomatis selama pelatihan dan prediksi model berlangsung.

Modeling and Training

Regresi Random Forest digunakan dalam penelitian ini; ini adalah algoritme pembelajaran ensemble berbasis pohon keputusan yang bekerja dengan menciptakan banyak keputusan secara acak dan kemudian menggabungkan semua riskor keputusan tersebut untuk menghasilkan prediksi akhir, untuk mengimplementasikan model ini, model itu sendiri dikonfigurasi untuk menggunakan 100 estimator, yang berarti bahwa model itu sendiri akan membentuk total 100 pohon keputusan yang berbeda. Prediksi akhir adalah rata-rata dari semua prediksi pohon lainnya, sehingga prediksi ini jauh lebih stabil dan akurat daripada satu pohon decision tree.

Untuk mempertahankan consistensi datanya dan memastikan reprodusibilitas selama pelatihan dan pengujian, nilai parameter *random_state* digunakan adalah 42. Ini penting agar metode pembagian data, bootstrap sampling, dan pemilihan fitur acak di setiap pohon tetap identik setiap kali model dijalankan ulang, terutama selama eksperimen. Dataset dibagi menjadi dua bagian sebesar 75% dan 25%, dan fungsi *train_test_split* dari *Scikit-Learn* digunakan. Hal ini memungkinkan model untuk mempelajari pola dari sebagian besar data, sementara masih dapat dengan obyektif menguji performa pada data yang secara tidak acak sebenarnya belum dilihat. Selain itu, ini juga harus dilakukan untuk mencegah data leakage, yang dapat meningkatkan bias prediksi.

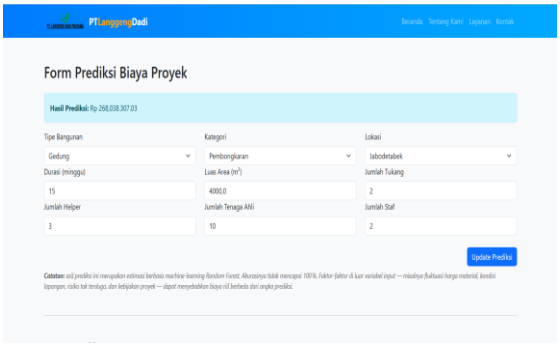
Seluruh proses pelatihan model dilakukan dengan satu alur yang mengintegrasikan keseluruhan proses tersebut menggunakan *Pipeline* dari *Scikit-Learn*. Pipeline dirancang untuk menyatukan seluruh tahapan pemrosesan ke alur yang berurutan. Pipeline terdiri dari dua komponen utama, yang pertama adalah pra-pemrosesan data. Pra-pemrosesan ini mencakup normalisasi fitur numerik menggunakan *MinMaxScaler.* Fungsi ini penting untuk memastikan seluruh fitur memiliki nilai yang seragam, yaitu di antara 0 hingga 1. Selain itu, fitur kategori dikonversi menjadi numerik dengan *one-hot encoding.* Hal ini diperlukan untuk memastikan bahwa model regresi dapat memproses data masukan yang diberikan dengan benar. Komponen kedua dari pipeline adalah regresi pohon acak. Pada tahap ini, model dilatih, dan nilai kontrak yang ada diprediksi pada fitur input. Pelatihan menggunakan data yang disiapkan dalam pipa dilakukan pada waktu

dan sumber daya yang relatif singkat. Karena random forest adalah algoritma yang dapat diparalelisasi dan berjumlah fitur input yang lebih kecil, pelatihan dapat diselesaikan dalam satu menit atau kurang pada perangkat keras biasa.

Kelebihan ini menunjukkan bahwa model tersebut sangat cocok untuk diintegrasikan dalam sistem prediksi. Algoritma tidak memerlukan sumber daya komputasi yang memiliki spesifikasi tinggi. Model ini dapat diambil real time dan diterapkan dalam jejaring, sehingga pengguna dapat memasukkan data proyek dan menerima prediksi nilai kontrak sesegera mungkin. Model prediksi tidak akan tertunda dalam sisi server karena komputasi ringannya

Development

Sistem prediksi biaya proyek konstruksi pada tahap pengembangan dibuat dengan Flask yang terhubung dengan model Random Forest Regressor yang dihasilkan. Dalam penggunaan, model berdasarkan data input pengguna dalam prediksi nilai kontrak proyek konstruksi. Data proyek dapat dimasukkan oleh pengguna menggunakan form prediksi biaya konstruksi yang ada di halaman sistem berbasis web ini. Berikut adalah gambaran sistem:



Gambar 2. Tampilan Form Prediksi Biaya Proyek Web Aplikasi

Pada gambar 2 ditunjukkan form tempat pengguna mengisi informasi-informasi tersebut, contohnya jenis project, kategori project, lokasi project, durasi pengerjaan, jumlah tukang, helper, tenaga ahli, staf, dan luas bangunan. Estimasi biaya akan langsung muncul setelah informasi dikirim.

Sebagai contoh, jika data-project pengguna mendefinisikan project dengan durasi sebanyak 15 minggu, luas bangunan 4000m 2, dan total tenaga kerja 17 orang, maka sistem akan menghitung harga kontrak-sepakat dari project tersebut senilai 268.038.307,03 rupiah. Angka tersebut dihasilkan dari rata-rata output logaritmik dari semua pohon dalam model *Random Forest*, kembalikan ke skala asli menggunakan fungsi expm1. Selain itu, kecuali memberikan hasil, sistem juga menyimpan seluruh informasi dalam database, dan menampilkannya ke dalam tabel histori di bawah form. Ini tentu bahwa dapat membantu pengguna dalam pelacakan dan perbandingan kalkulasi project sebelum, sistem tulisan ini mampu memberi pengalaman pengguna yang baik, dan meredakan proses perencanaan budget project dengan cepat dan keterjangkauan data.

Visualisasi Evaluasi Model

Evaluasi model dilakukan untuk menentukan bagaimana algoritma Random Forest Regressor dapat memprediksi nilai kontrak proyek konstruksi secara akurat, Evaluasi ini menggunakan empat metrik utama:

Diambil data uji

i	y_i	$\bar{y}\, i$
1	6.000.000.000	5.100.000.000
2	5.800.000.000	5.000.000.000
3	5.500.000.000	5.800.000.000
4	6.200.000.000	5.700.000.000

5	5.900.000.000	5.200.000.000

Mean Absolute Error (MAE)

menghitung perbedaan absolut antara nilai aktual dan nilai prediksi.

MAE = $\frac{1}{n}\sum_{i=1}^{n} |y_i - \bar{\hat{y}}\, i|$

yi = nilai aktual

$\bar{\hat{y}}\, i$ = nilai prediksi

n = jumlah data

i	$y_i - \bar{\hat{y}}\ i$
1	900.000.000
2	800.000.000
3	300.000.000
4	500.000.000
5	700.000.000

MAE = $\frac{900 + 800 + 300 + 500 + 700}{5} = \frac{3.200.000.000}{5} = 640.000.000$

Root Mean Squared Error (RMSE)

RMSE = $\sqrt{\sum_{i=1}^{n}(y_i - \bar{\hat{y}}\ i)^{\ 2}}$

Perhitungan

i	$y_i - \bar{\hat{y}}\ i$
1	8.1×10^{17}
2	6.4×10^{17}
3	9.0×10^{16}
4	2.5×10^{17}
5	4.9×10^{17}

Total = 2.229×10^{18}

RMSE = $\sqrt{\frac{2.229e18}{5}} = \sqrt{4.458e17} \approx 667.816.000$

R-squared (R^2)

$$R^2 = 1 - \frac{\sum (yi - \bar{\hat{y}}\, i)^2}{\sum (yi - \bar{y})^2}$$

Hitung rata-rata nilai aktual :

$\bar{y}$= $\frac{6.0 + 5.8 + 5.5 + 6.2 + 5.9}{5}$ = 5.88 milyar

i	$yi - \bar{y}$
1	0.0144
2	0.0004
3	0.1444
4	0.1024
5	0.0004

$\sum = (yi - \bar{y})^2 = 0.262 \times 10^{18}$

$$R^2 = 1 - \frac{2.229e18}{2.62e17} = 1 - 8.51 \approx -7.51$$

R^2 negatif menunjukkan model buruk karena prediksi lebih buruk dari rata-rata (bisa terjadi jika error sangat besar atau data terlalu sedikit).

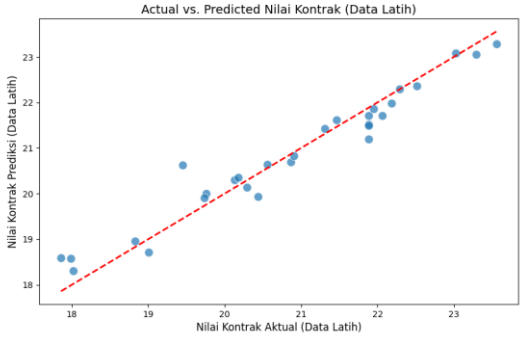
Mean Absolute Percentage Error (MAPE)

MAPE = $\frac{100\%}{n}\sum_{i=1}^{n} |\frac{yi - \hat{y}i}{yi}|$

i	%error
1	15%
2	13.79%
3	5.45%
4	8.06%
5	11.86%

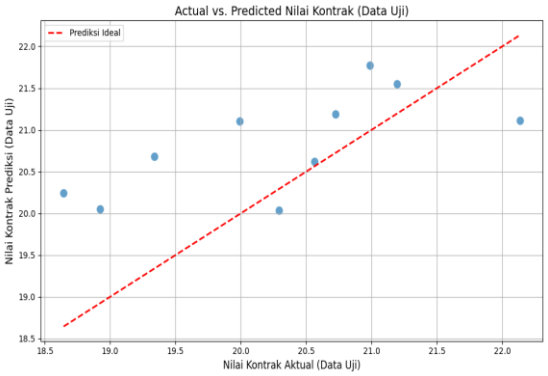
MAPE = $\frac{15 + 13.79 + 5.45 + 8.06 + 11.86}{5}$ = 10.83%

Untuk mendukung hasil evaluasi numerik, dilakukan visualisasi terhadap data latih dan data uji untuk memahami sejauh mana prediksi model mendekati nilai aktual, ada tiga jenis visualisasi digunakan:



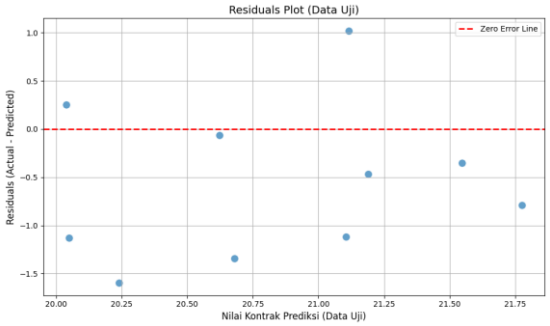
Gambar 3. Plot aktual vs. prediksi pada data latih

Pertama pada Gambar 3 Grafik tersebut secara jelas menggambarkan hubungan antara nilai aktual dan nilai prediksi pada data latih. Sebagian besar titik pada grafik cenderung membentuk pola yang mengikuti garis prediksi ideal atau garis diagonal, yang berarti bahwa model telah belajar dari pola dalam data pelatihan dengan baik. Observasi ini konsisten dengan hasil evaluasi kuantitatif, dengan R^2 = -0.262 dan MAPE = 10,83% semuanya dekat dengan nilai sesungguhnya.



Gambar 4. Plot aktual vs. prediksi pada data uji

Kedua pada Gambar 4 Hasil berikut adalah grafik Aktual vs. Prediksi untuk data uji. Dibandingkan dengan data pelatihan, titik tersebar juga cenderung lebih menyebar dan lebih banyak bercerai dari garis ideal, terutama untuk proyek-proyek dengan kontrak tinggi. Alasan di balik pola ini adalah model kesulitan dalam identifikasi pola dari data yang tidak diketahui sebelumnya. Ini didukung oleh hasil evaluasi data uji, yang terjadi penurunan kinerja pada data ini. Dengan R^2 sebesar -0.262 dan MAPE 10,83% , prediksi model kurang akurat.



Gambar 5. Residuals plot pada data uji

Ketiga Gambar 5 memperlihatkan residuals plot pada data uji, di mana residual adalah pemisahan antara nilai aktual dan nilai prediksi. Dalam hal ini, kebanyakan titik residual berada di bawah garis nol, yang berarti model cenderung meremehkan nilai kontrak, atau memiliki bias negatif. Bahkan jika distribusi residual tidak acak dan tidak mendukung pola, ini menunjukkan bahwa tidak ada heteroskedastisitas yang jelas, bisa negatifnya bisa berarti akurasi model masih belum optimal, terutama untuk proyek-proyek berskala besar. Dengan demikian, hal ini membuat catatan penting untuk mengingat ketika ada kemungkinan model bisa diperbaiki dengan menambah fitur atau tuning parameter yang lebih halus.

KESIMPULAN DAN SARAN

Penelitian ini telah berhasil memajukan sistem prediksi biaya proyek konstruksi berbasis web menggunakan algoritma Random Forest Regressor, yang dapat membantu pengguna memperkirakan biaya proyek sesuai input seperti jenis bangunan, lokasi, durasi, jumlah pekerja, dan luas bangunan. Berdasarkan hasil evaluasi, model memiliki nilai Mean Absolute Error sebesar Rp 640.000.000 dan Mean Absolute Percentage Error sebesar 10,83%, yang sudah cukup baik, meski nilai R-squared sebesar -0.262 menandakan

bahwa model ini belum merupakan model yang optimal dalam memprediksi data baru. Oleh karena itu, model ini masih dapat memperbaiki result dengan sampel data pada proyek yang lebih besar dan lebih variatif, disertai dengan penyempurnaan model ditambah dengan fitur baru dan beberapa pengaturan parameter. Selain itu, validasi hasil dengan standar proyek yang ada di lapangan juga masih perlu dilakukan agar hasil prediksi lebih akurat dan dapat dijadikan acuan ke pemborongannya. Untuk perkembangan sistem lebih lanjut agar terintegrasi dengan manajemen proyek sistem keseluruhan juga akan meningkatkan keuntungan dari sistem ini di industri konstruksi.

DAFTAR PUSTAKA

[1] [1] N. Putu, I. Yuliana, N. Kadek, and S. Ebtha, "*M Anajemen R Isiko E Stimasi B Iaya P Ada T Ahap P Erencanaan P Royek K Onstruksi R Isk M Anagement F or C Ost E Stimation a T T He P Lanning*," pp. 92–98, 2024.

[2] K. Krisnandi and H. Agung, "*Implementasi Algoritma Genetika Untuk Memprediksi Waktu Dan Biaya Pengerjaan Proyek Konstruksi*," J. Ilm. FIFO, vol. 9, no. 2, p. 90, 2017, doi: *10.22441/fifo.2017.v9i2.001*.

[3] P. A. Eman, L. Elisabeth, and F. Jansen, "*Estimasi Biaya Konstruksi Menggunakan Metode Parameter Pada Proyek Pemeliharaan Berkala Jalan Di Kota Manado*," J. Ilm. Media Eng., vol. 8, no. 2, p. 1033, 2018.

[4] K. N. Knn, A. Ernawati, and Z. Sitorus, "*Penerapan Data Mining Untuk Klasifikasi Penduduk Miskin Di Kabupaten Labuhanbatu Menggunakan Random Forest (RF) Dan*," vol. 6, no. 1, pp. 23–29, 2025, doi: *10.47065/bit.v5i2.1783*.

[5] G. A. B. Suryanegara, Adiwijaya, and M. D. Purbolaksono, "*Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi*," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 1, no. 10, pp. 114–122, 2021.

[6] S. Kusworo, N. A. Santoso, and R. D. Kurniawan, "*Prediksi Nilai Akhir Semester Siswa Menggunakan Algoritma Random Forest*," no. 2, 2024.

[7] K. All, D. Not, and L. Regression, "*Prediksi Harga Smartphone Berdasarkan Fitur Smartphone Dengan Random Forest Regression*," vol. 50, no. x, pp. 1–13, 2006.

[8] Farhanuddin, Sarah Ennola Karina Sihombing, and Yahfizham, "*Komparasi Multiple Linear Regression dan Random Forest Regression Dalam Memprediksi Anggaran Biaya Manajemen Proyek Sistem Informasi*," J. Comput. Digit. Bus., vol. 3, no. 2, pp. 86–97, 2024, *doi: 10.56427/jcbd.v3i2.408*.

[9] E. S. Lestari and I. Astuti, "*Penerapan Random Forest Regression Untuk Memprediksi Harga Jual Rumah Dan Cosine Similarity Untuk Rekomendasi Rumah Pada Provinsi Jawa Barat*," *J. Ilm. FIFO*, vol. 14, no. 2, p. 131, 2022, doi: *10.22441/fifo.2022.v14i2.003*.

[10] T. Pustaka, "*Penerapan Algoritma Random Forest Untuk Prediksi Pejualan Dan Persedian Produk Pada Toko Frozen Food Anisa*," pp. 843–848, 2025.

[11] I. Romli, "*Penerapan Data Mining Menggunakan Algoritma K-Means Untuk Klasifikasi Penyakit Ispa*," Indones. J. Bus. Intell., vol. 4, no. 1, p. 10, 2021, doi: *10.21927/ijubi.v4i1.1727*.

[12] W. A. Naseer, S. Sarwido, and B. B. Wahono, "*Gradient Boosting Optimization with Pruning Technique for Prediction of Bmt Al-hikmah*

Permata Customer Data," Jinteks, vol. 6, no. 3, pp. 719–727, 2024.
[13] M. A. I. N. R. Said, Mohammad Reza Faisal, Dwi Kartini, Irwan Budiman, and Triando Hamonangan Saragih, "*Analisis Perbandingan Metode Harmonic Meandan Local Mean VectorDalam Penyeleksian Tetangga Pada Algoritma KNN*," J. Sains dan Inform., vol. 9, no. 2, pp. 127–135, 2023, doi: *10.34128/jsi.v9i2.376.*
[14] T. Informatika *et al.*, "*Tisanalisis Prediksi Penjualan Tisu Menggunakan Regresi Linear*," vol. 13, no. 2, pp. 739–746, 2025.
[15] M. S. Efendi and A. K. Zyen, "*Penerapan Algoritma Random Forest Untuk Prediksi Penjualan Dan Sistem Persediaan Produk*," vol. 5, no. 1, pp. 12–20, 2024, doi: *10.30865/resolusi.v5i1.2149.*
[16] E. H. Y. K. Eben Haezer and N. Setiyawati, "*Pembangunan Aplikasi Virtual Inventory System (Vis) Berbasis Web Menggunakan Flask Framework*," J. Mnemon., vol. 5, no. 2, pp. 128–135, 2022, doi: 10.36040/mnemonic.v5i2.4799.
[17] "Journal of Artificial Intelligence and Digital Business (RIGGS) *Prediksi Harga Properti Di Indonesia Menggunakan Algoritma Random*," vol. 4, no. 1, pp. 43–49, 2025.
[18] I. Nabillah and I. Ranggadara, "*Mean Absolute Percentage Error untuk Evaluasi Hasil Prediksi Komoditas Laut*," JOINS (Journal Inf. Syst., vol. 5, no. 2, pp. 250–255, 2020, doi: *10.33633/joins.v5i2.3900.*
[19] M. H. Al Mustofa and Nurmalitasari, "*Analisis Harga Kuota Internet Menggunakan Metode Multiple Regresi Linear*," pp. 319–325, 2023.