BAB II

LANDASAN TEORI

2.1. Knowledge Discovery in Databases

Knowledge Discovery in Databases (KDD) merupakan proses sistematis untuk menemukan pola atau pengetahuan yang valid, baru, bermanfaat, dan dapat dipahami dari sekumpulan data. KDD tidak hanya berfokus pada pengoperasian algoritma. Proses ini mencakup pemahaman tujuan, pemilihan data, pembersihan, transformasi, pemodelan, evaluasi, dan penafsiran hasil menjadi pengetahuan yang dapat digunakan. Definisi tersebut menempatkan kualitas data dan ketepatan interpretasi sebagai bagian yang tidak terpisahkan dari analisis (U. Fayyad, G. Piatetsky-Shapiro, 1996).

KDD bersifat interaktif dan berulang. Hasil pada tahap pemodelan dapat menunjukkan bahwa data perlu dibersihkan kembali, variabel perlu ditinjau ulang, atau teknik transformasi perlu diperbaiki. Model proses KDD membantu peneliti mendokumentasikan hubungan antartahap dan menjaga keterlacakan keputusan analitis (KURGAN et al., 2006). Dengan demikian, keluaran KDD bukan hanya nilai prediksi, tetapi juga pengetahuan yang dapat dijelaskan sesuai tujuan dan batas penelitian.

Dalam penelitian ini, KDD digunakan sebagai kerangka pengolahan data pelanggan listrik prabayar. Setiap observasi mewakili satu pelanggan. Lima variabel prediktor terdiri atas MERK_KWH, TARIF, DAYA_VA, NAMA_KECAMATAN, dan NAMA_KELURAHAN. Variabel target adalah

KONSUMSI_RATA_RATA_TARGET. Seluruh proses diarahkan untuk membangun prediksi konsumsi listrik rata-rata antarpelanggan, bukan untuk meramalkan deret waktu dan bukan untuk menyimpulkan hubungan sebab akibat.

2.2. Tahapan Knowledge Discovery in Databases

Tahapan KDD dalam penelitian ini mengikuti alur seleksi data, prapemrosesan, transformasi, data mining, serta evaluasi dan interpretasi. Urutan tersebut memberikan kerangka kerja yang logis, tetapi penerapannya tidak selalu berjalan satu arah. Peneliti dapat kembali ke tahap sebelumnya apabila menemukan masalah kualitas data atau hasil model yang tidak memadai (U. Fayyad, G. Piatetsky-Shapiro, 1996).

2.2.1. Seleksi Data

Seleksi data bertujuan menentukan observasi dan variabel yang relevan dengan tujuan penelitian. Data yang dipilih mencakup pelanggan listrik prabayar pada wilayah penelitian dengan informasi prediktor dan target yang diperlukan. ID_RECORD hanya digunakan untuk pelacakan data dan tidak dimasukkan sebagai prediktor karena tidak memiliki makna prediktif substantif.

Variabel rerata_rpptl dan jam_nyala tidak digunakan sebagai prediktor karena berisiko membawa informasi yang berkaitan langsung dengan pembentukan target. Variabel jenislayanan juga tidak digunakan karena seluruh data akhir berada pada satu kategori prabayar sehingga tidak memiliki variasi. Identitas pribadi pelanggan dikeluarkan untuk menjaga kerahasiaan. Keputusan

seleksi tersebut menjaga kesesuaian data dengan tujuan prediksi dan mengurangi risiko kebocoran informasi.

2.2.2. Prapemrosesan Data

Prapemrosesan dilakukan untuk memastikan data memiliki kualitas yang memadai sebelum pemodelan. Pemeriksaan mencakup nama kolom, tipe data, nilai kosong, keunikan ID_RECORD, baris identik, nilai target negatif, kewajaran DAYA_VA, keseragaman kategori, dan nilai ekstrem. Nilai ekstrem tidak langsung dihapus karena dapat merepresentasikan variasi konsumsi yang sah. Perubahan hanya dilakukan apabila terdapat bukti kesalahan pencatatan atau alasan substantif yang terdokumentasi.

2.2.3. Transformasi Data

Transformasi mengubah data ke dalam representasi yang dapat diproses algoritma. Empat prediktor kategorikal diubah melalui one-hot encoding agar setiap kategori direpresentasikan tanpa membentuk urutan semu. DAYA_VA tetap dipertahankan sebagai variabel numerik dan tidak memerlukan standardisasi untuk Random Forest Regression. Pemilihan teknik pengodean perlu mempertimbangkan karakteristik kategori dan algoritma yang digunakan (Hancock & Khoshgoftaar, 2020) .

2.2.4. Evaluasi dan Interpretasi

Kinerja model diukur menggunakan *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE), *koefisien determinasi* (R^2), dan *Symmetric Mean Absolute Percentage Error* (sMAPE). RMSE digunakan sebagai metrik utama karena memberikan penalti lebih besar pada kesalahan besar dan tetap memiliki satuan yang sama dengan target. *Random Forest* juga dibandingkan dengan baseline berbasis rata-rata agar manfaat prediktifnya dapat dinilai secara objektif.

Tahap interpretasi menggunakan permutation importance. Teknik ini mengukur perubahan RMSE setelah nilai suatu prediktor diacak. Peningkatan RMSE yang lebih besar menunjukkan bahwa model lebih bergantung pada prediktor tersebut. Hasilnya menjelaskan kontribusi prediktif dalam model, bukan hubungan sebab akibat.

2.3. Data Mining

Data mining merupakan tahap dalam KDD yang menerapkan algoritma untuk menemukan pola, hubungan, atau model dari data yang telah dipersiapkan. Tugas data mining dapat berupa klasifikasi, regresi, pengelompokan, deteksi anomali, dan penemuan aturan asosiasi. Pemilihan tugas ditentukan oleh tujuan analisis dan bentuk variabel target (U. Fayyad, G. Piatetsky-Shapiro, 1996).

KDD dan data mining tidak boleh digunakan sebagai istilah yang sepenuhnya sama. KDD mencakup keseluruhan proses pengubahan data menjadi pengetahuan, sedangkan data mining berfokus pada penerapan algoritma. Dalam

penelitian ini, keseluruhan rangkaian pengolahan data mengikuti KDD, sementara penerapan *Random Forest Regression* menjadi kegiatan data mining untuk menyelesaikan tugas regresi.

Hasil data mining perlu dievaluasi terhadap data yang tidak digunakan untuk melatih model. Nilai kinerja pada data training tidak cukup untuk menyatakan bahwa pola dapat diterapkan pada pelanggan lain. Oleh karena itu, pemisahan data, validasi silang, baseline, dan pengujian akhir menjadi bagian penting untuk menjaga validitas hasil prediksi.

2.4. Machine Learning

Machine learning merupakan bidang yang mengembangkan metode komputasi agar sistem dapat mempelajari pola dari data dan menggunakan pola tersebut untuk menghasilkan prediksi pada data baru. Bidang ini berada pada pertemuan ilmu komputer dan statistika. Kinerjanya dipengaruhi oleh kualitas data, representasi variabel, algoritma, prosedur pelatihan, dan metode evaluasi (Mitchell, 2015).

Machine learning dapat dikelompokkan menjadi *supervised learning*, *unsupervised learning*, dan *reinforcement learning*. Penelitian ini termasuk supervised learning karena setiap observasi memiliki target konsumsi listrik rata-rata yang diketahui. Model mempelajari pemetaan dari lima prediktor menuju target numerik kontinu.

Tujuan pembelajaran tidak berhenti pada pencocokan data training. Model harus memiliki kemampuan generalisasi, yaitu mempertahankan kinerja ketika

menerima data yang tidak digunakan pada pelatihan. Perbedaan besar antara kinerja data training dan data testing dapat menunjukkan overfitting. Sebaliknya, kinerja yang sama-sama rendah dapat menunjukkan bahwa model belum menangkap pola yang relevan atau informasi prediktornya terbatas (Trevor Hastie , Robert Tibshirani, 2009).

Dalam penelitian ini, hasil *machine learning* ditafsirkan sebagai hubungan prediktif. Model dapat menunjukkan bahwa kombinasi karakteristik pelanggan membantu memperkirakan konsumsi rata-rata, tetapi hasil tersebut tidak membuktikan bahwa suatu karakteristik menyebabkan perubahan konsumsi. Batas interpretasi ini harus dipertahankan pada pembahasan dan kesimpulan.

2.5. Regresi

Regresi merupakan metode untuk memodelkan hubungan antara satu atau lebih variabel prediktor dan target numerik kontinu. Secara umum, hubungan tersebut dapat ditulis sebagai:

$$y = f(X) + \varepsilon$$

Pada persamaan tersebut, y merupakan variabel target, X merupakan kumpulan variabel prediktor, f(X) merupakan pola hubungan yang dipelajari model, dan ε merupakan bagian kesalahan yang tidak dapat dijelaskan oleh model. Dalam penelitian ini, y adalah KONSUMSI_RATA_RATA_TARGET dan X terdiri atas MERK_KWH, TARIF, DAYA_VA, NAMA_KECAMATAN, serta NAMA_KELURAHAN.

Regresi tidak selalu berbentuk linear. Regresi linear menetapkan bentuk hubungan tertentu melalui kombinasi koefisien, sedangkan metode berbasis pohon dapat mempelajari pola nonlinier dan interaksi antarvariabel tanpa menetapkan persamaan linear sejak awal. *Random Forest Regression* dipilih karena data memuat kombinasi prediktor numerik dan kategorikal serta memungkinkan hubungan yang kompleks (Jin et al., 2020).

Selisih antara nilai aktual dan nilai prediksi disebut residual. Residual menjadi dasar perhitungan metrik kesalahan. Model yang baik menghasilkan residual yang relatif kecil pada data testing dan memberikan kinerja yang lebih baik daripada baseline. Evaluasi tetap dilakukan dengan beberapa metrik karena setiap metrik menyoroti karakteristik kesalahan yang berbeda.

2.6. Random Forest Regression

Random Forest merupakan metode ensemble yang menggabungkan banyak pohon keputusan. Setiap pohon dibangun menggunakan sampel bootstrap dan hanya mempertimbangkan sebagian prediktor secara acak pada setiap pemisahan. Mekanisme tersebut menghasilkan keragaman antarpohon dan membantu mengurangi ketergantungan model pada satu pohon tertentu (Jin et al., 2020).

Pada tugas regresi, setiap pohon menghasilkan satu prediksi numerik. Prediksi akhir diperoleh dari rata-rata prediksi seluruh pohon. Secara matematis, prediksi *Random Forest Regression* dapat dinyatakan sebagai:

$$\hat{y}(x) = (1/B) \Sigma_{\beta=1}^{B} T_{\beta}(x)$$

Keterangan: ŷ(x) merupakan hasil prediksi akhir untuk data x, B merupakan jumlah pohon, dan $T_{\beta}(x)$ merupakan hasil prediksi pohon ke-b. Penggabungan prediksi membantu mengurangi varians dibandingkan satu pohon yang tumbuh dalam dan peka terhadap perubahan data.

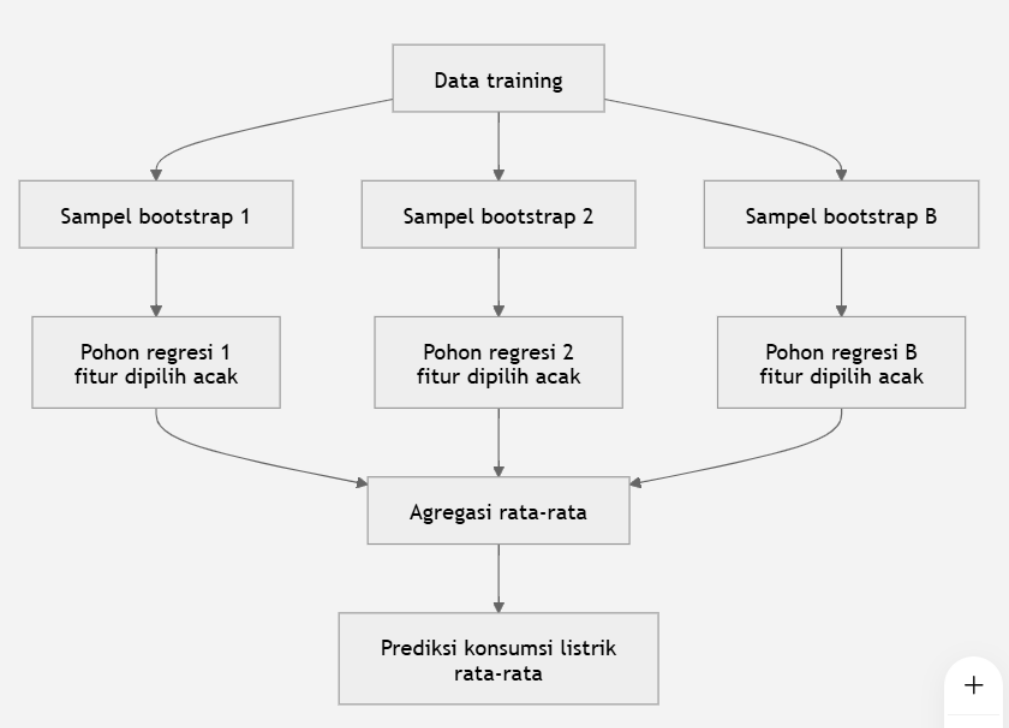
2.6.1. Prinsip Kerja Random Forest Regression

Proses dimulai dengan *bootstrap* sampling, yaitu pengambilan sampel dari data training dengan pengembalian. Setiap sampel digunakan untuk membangun satu pohon regresi. Pada setiap node, algoritma memilih kandidat pemisahan dari subset prediktor yang diambil secara acak. Pohon terus tumbuh sesuai batas hyperparameter, lalu prediksi seluruh pohon dirata-ratakan (Pedregosa et al., 2011).

2.6.2. Hyperparameter Random Forest Regression

Hyperparameter mengendalikan kompleksitas dan proses pembentukan model. *n_estimators* menentukan jumlah pohon. *max_depth* membatasi kedalaman pohon. *min_samples_split* menentukan jumlah minimum observasi untuk membagi suatu node. *min_samples_leaf* menentukan jumlah minimum observasi pada daun. *max_features* menentukan jumlah prediktor yang dipertimbangkan pada setiap pemisahan. Pengaruh setiap parameter perlu dinilai melalui validasi, bukan berdasarkan asumsi tunggal (Probst et al., 2019)

Penelitian ini menggunakan *GridSearchCV* pada data training untuk menguji kombinasi *hyperparameter* secara sistematis. *K-fold cross-validation* memberikan estimasi kinerja pada beberapa pembagian data training. Kombinasi terbaik dipilih berdasarkan RMSE validasi, kemudian model akhir dievaluasi satu kali pada data testing yang tidak digunakan selama pemilihan model.



Gambar 2.1 Alur Kerja Random Forest Regression

Gambar 2.1 menunjukkan bahwa data training digunakan untuk membentuk beberapa sampel *bootstrap*. Setiap sampel melatih satu pohon regresi dengan pemilihan fitur secara acak. Prediksi akhir diperoleh melalui rata-rata prediksi seluruh pohon. Alur ini berlangsung di dalam *pipeline* setelah data kategorikal ditransformasikan dan DAYA_VA dipertahankan sebagai variabel numerik.

2.7. Prapemrosesan Data

Prapemrosesan data merupakan tahap untuk memastikan bahwa data memiliki kualitas, struktur, dan format yang sesuai sebelum pemodelan. Tahap ini harus ditetapkan secara transparan karena keputusan pembersihan dan transformasi dapat memengaruhi kinerja serta interpretasi model. Dalam penelitian ini, prapemrosesan juga dirancang untuk mencegah data leakage.

2.7.1. Pemeriksaan Data Kosong dan Duplikat

Nilai kosong diperiksa pada seluruh prediktor dan target. Penanganannya ditentukan berdasarkan jumlah, pola, dan arti variabel. Baris identik tidak langsung dianggap duplikat yang harus dihapus karena pelanggan berbeda dapat memiliki karakteristik dan nilai konsumsi yang sama. ID_RECORD digunakan untuk membedakan observasi selama audit, tetapi tidak digunakan dalam pelatihan.

2.7.2. Pemeriksaan Tipe dan Nilai Data

Prediktor kategorikal harus memiliki penulisan yang konsisten, sedangkan DAYA_VA dan target harus tersimpan sebagai nilai numerik. Pemeriksaan juga mencakup nilai negatif, nilai di luar kewajaran, dan kategori berfrekuensi kecil. Setiap koreksi atau penghapusan harus didukung alasan yang dapat ditelusuri dan tidak boleh dilakukan hanya untuk memperbaiki nilai evaluasi.

2.7.3. Pengodean Variabel Kategorikal

MERK_KWH, TARIF, NAMA_KECAMATAN, dan NAMA_KELURAHAN merupakan prediktor kategorikal. One-hot encoding mengubah setiap kategori menjadi kolom biner tanpa memberikan peringkat buatan. Opsi *handle_unknown='ignore'* memungkinkan *pipeline* memproses kategori pada data testing yang tidak muncul dalam bagian training tanpa menghentikan proses prediksi. Representasi kategori perlu dipilih secara sadar karena dapat memengaruhi proses pembelajaran model (Hancock & Khoshgoftaar, 2020).

2.7.4. Pembagian Data dan Pencegahan Data Leakage

Data dibagi menjadi data training dan data testing sebelum *encoder* mempelajari kategori. Data training digunakan untuk prapemrosesan, validasi silang, penyetelan *hyperparameter*, dan pelatihan. Data testing dipertahankan sebagai holdout untuk evaluasi akhir. Pemisahan tersebut mencegah informasi data testing memengaruhi proses pemilihan model.

Data *leakage* terjadi ketika informasi yang tidak seharusnya tersedia pada tahap pelatihan ikut digunakan untuk membangun model sehingga nilai evaluasi menjadi terlalu optimistis. Kebocoran dapat berasal dari target, data testing, transformasi yang dipelajari dari seluruh data, atau variabel turunan yang membawa informasi target (Kaufman et al., 2012). Karena itu, rerata_rpptl dan jam_nyala dikeluarkan, sedangkan encoder ditempatkan di dalam pipeline.

2.7.5. Pipeline Prapemrosesan

Pipeline menyatukan *OneHotEncoder, ColumnTransformer,* dan *Random Forest Regression*. *ColumnTransformer* menerapkan pengodean pada empat prediktor kategorikal dan meneruskan DAYA_VA sebagai prediktor numerik. Konfigurasi ini harus memastikan DAYA_VA tidak terhapus oleh pengaturan remainder. *Pipeline* juga menjamin bahwa transformasi dipelajari ulang hanya dari bagian training pada setiap fold validasi (Pedregosa et al., 2011).

2.8. Evaluasi Model Prediksi

Evaluasi model mengukur perbedaan antara nilai aktual dan nilai prediksi pada data yang tidak digunakan untuk menyesuaikan model. Penelitian ini menggunakan MAE, MSE, RMSE, R^2, dan sMAPE. Penggunaan beberapa metrik diperlukan karena satu ukuran tidak selalu menggambarkan seluruh karakteristik kesalahan regresi (Chicco et al., 2021).

2.8.1. Mean Absolute Error

Mean Absolute Error (MAE) menghitung rata-rata nilai absolut selisih antara nilai aktual dan prediksi:

$$MAE = (1/n)\ \Sigma_{i=\ 1}^{\ n}\ |y_i - \hat{y}_i|$$

MAE memiliki satuan yang sama dengan variabel target. Nilai yang lebih kecil menunjukkan rata-rata kesalahan absolut yang lebih rendah. MAE memperlakukan setiap kesalahan secara proporsional sehingga relatif mudah ditafsirkan.

2.8.2. Mean Squared Error

Mean Squared Error (MSE) menghitung rata-rata kuadrat selisih nilai aktual dan prediksi:

$$MSE = (1/n) \Sigma_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Pengkuadratan memberikan penalti lebih besar pada kesalahan yang besar. Satuan MSE merupakan kuadrat dari satuan target sehingga nilainya tidak dapat ditafsirkan secara langsung dalam satuan konsumsi asli.

2.8.3. Root Mean Squared Error

Root Mean Squared Error (RMSE) merupakan akar kuadrat MSE:

$$RMSE = \sqrt{[(1/n) \Sigma_{i=1}^{n} (y_i - \hat{y}_i)^2]}$$

RMSE kembali memiliki satuan yang sama dengan target dan tetap sensitif terhadap kesalahan besar. Penelitian ini menetapkan RMSE sebagai metrik utama untuk validasi dan permutation importance. Nilai RMSE yang lebih kecil menunjukkan kinerja prediksi yang lebih baik.

2.8.4. Koefisien Determinasi

Koefisien determinasi atau R^2 mengukur kemampuan model menjelaskan variasi target pada data evaluasi:

$$R^2 = 1 - [\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2 / \Sigma_{i=1}^{n}(y_i - \bar{y})^2]$$

Nilai R^2 yang mendekati satu menunjukkan kemampuan prediktif yang lebih baik. Nilai nol menunjukkan kinerja yang setara dengan prediksi rata-rata

pada data evaluasi. R^2 dapat bernilai negatif ketika model lebih buruk daripada tolok ukur rata-rata. Oleh karena itu, nilai negatif tidak boleh dihapus atau dipaksa menjadi nol.

2.8.5. Symmetric Mean Absolute Percentage Error

Symmetric Mean Absolute Percentage Error (sMAPE) menyatakan kesalahan relatif dalam bentuk persentase:

$$sMAPE = (100\%/n) \Sigma_{i=1}^{n} [2|y_i - \hat{y}_i| / (|y_i| + |\hat{y}_i|)]$$

Dengan rumus tersebut, rentang teoritis sMAPE adalah 0 sampai 200 persen. Nilai yang lebih kecil menunjukkan kesalahan relatif yang lebih rendah. Ketika nilai aktual dan prediksi sama-sama nol, kontribusi observasi ditetapkan nol untuk mencegah pembagian dengan nol. Ukuran berbasis persentase perlu ditafsirkan hati-hati ketika nilai aktual mendekati nol (Hyndman & Koehler, 2006).

2.8.6. Validasi Silang dan Perbandingan Baseline

K-fold cross-validation membagi data training menjadi K bagian. Pada setiap putaran, satu bagian digunakan sebagai validation dan bagian lainnya digunakan untuk melatih model. Proses diulang sampai setiap bagian menjadi validation. Nilai kinerja dirata-ratakan untuk menilai stabilitas kombinasi hyperparameter (Arlot & Celisse, 2010).

Random Forest dibandingkan dengan *DummyRegressor* yang selalu memprediksi rata-rata target pada data training. Kedua model dievaluasi pada

data testing yang sama. *Random Forest* dinilai memiliki manfaat prediktif apabila menghasilkan RMSE, MAE, MSE, dan sMAPE yang lebih rendah serta R^2 yang lebih baik daripada baseline.

2.9. Permutation Importance

Permutation importance merupakan teknik interpretasi model yang mengukur perubahan kinerja setelah nilai suatu prediktor diacak. Pengacakan memutus hubungan prediktif antara variabel tersebut dan target, sementara variabel lain dipertahankan. Jika kinerja menurun secara besar, model dinilai lebih bergantung pada prediktor yang diacak (Jin et al., 2020).

Dalam penelitian ini, model terbaik terlebih dahulu dievaluasi pada data asli untuk memperoleh RMSE acuan. Setiap prediktor kemudian diacak dan RMSE dihitung kembali. Importance prediktor j dapat dituliskan sebagai:

$$Importance_j = RMSE\ setelah\ permutasi_j - RMSE\ acuan$$

Nilai positif yang lebih besar menunjukkan peningkatan kesalahan yang lebih besar setelah pengacakan. Pengacakan dilakukan sebanyak 20 kali untuk memperoleh nilai rata-rata dan variasi *importance.* Pengulangan membantu mengurangi ketergantungan hasil pada satu susunan acak.

Konsep *permutation-based importance* telah digunakan dalam *Random Forest* (Jin et al., 2020). (Altmann et al., 2010) mengembangkan prosedur koreksi yang menggunakan permutasi target untuk menilai signifikansi. Penelitian ini hanya menggunakan permutation importance pada scikit-learn dan tidak menghitung *p-value*. Oleh karena itu, metode yang digunakan tidak disebut PIMP.

Permutation importance memiliki keterbatasan ketika prediktor saling berkorelasi atau membawa informasi yang tumpang tindih. Pengacakan satu prediktor dapat menghasilkan penurunan kinerja yang kecil karena informasi serupa masih tersedia pada prediktor lain. Nilai importance juga menjelaskan ketergantungan model dalam data penelitian dan tidak menunjukkan besar pengaruh kausal.

2.10. Penelitian Terdahulu

Penelitian terdahulu digunakan untuk menempatkan penelitian dalam perkembangan penerapan *Random Forest Regression*. Kajian mencakup prediksi konsumsi energi dan aplikasi regresi pada domain lain.

Tabel 2.1 Penelitian Terdahulu

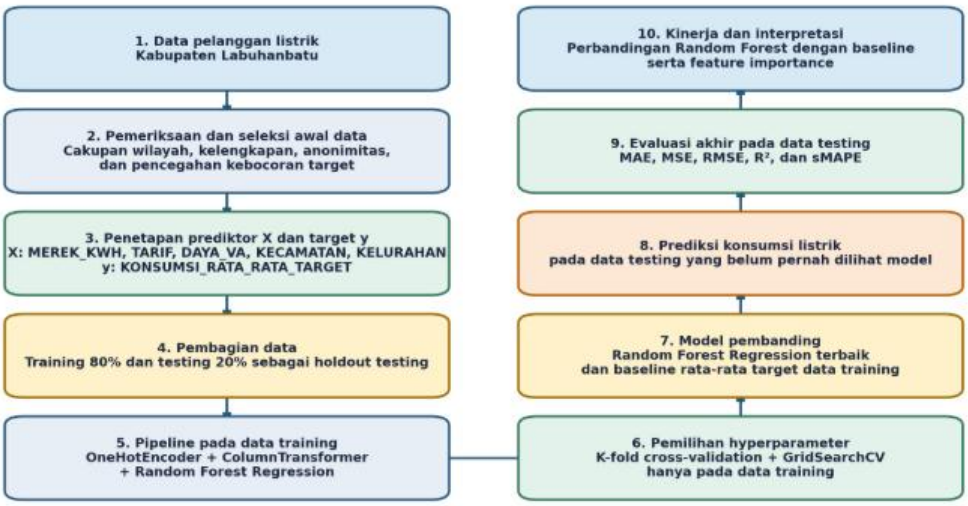
No	Peneliti	Judul	Tahun	Data dan Metode di gunakan	Hasil Penelitian
1	Kamilah, N. N., dan Rahman, A. INOVTEK Polbeng-SI,	*Optimization of Household Energy Use Prediction Using Random Forest with Genetic Algorithm Feature Selection*	2025	Data konsumsi listrik rumah tangga; Random Forest Regression dan seleksi fitur berbasis Genetic Algorithm.	Random Forest Regression dapat digunakan untuk memprediksi konsumsi energi rumah tangga. Seleksi fitur membantu menyederhanakan prediktor, tetapi pengaruhnya terhadap kinerja model perlu dinilai melalui beberapa metrik.

No	Peneliti	Judul	Tahun	Data dan Metode di gunakan	Hasil Penelitian
2	Syafei, R. M., Nikmah, T. L., Anisa, D. N., dan Kharisma, S. N. JOISER,	*Optimization of Energy Consumption Prediction with Random Forest Regressor and XGBoost Feature Importance*	2026	Data konsumsi energi yang memuat variabel waktu, cuaca, karakteristik bangunan, dan penggunaan energi; Random Forest Regressor dan XGBoost	Random Forest Regressor dapat diterapkan untuk memprediksi konsumsi energi. Pemilihan fitur mendukung identifikasi variabel yang relevan dan penyederhanaan model.
3	Soraya, N. S., dan Hendry AITI, (Soraya & Hendry, 2023)	*Komparasi Linear Regression, Random Forest Regression, dan Multilayer Perceptron Regression untuk Prediksi Tren Musik TikTok*	2023	Data fitur audio dan popularitas musik TikTok; Linear Regression, Random Forest Regression, dan MLP Regression.	Hasil perbandingan menunjukkan bahwa Random Forest Regression tidak selalu menjadi model terbaik. Pemilihan algoritma perlu mempertimbangkan pola dan karakteristik data.
4	Listyaningrum, R., Purwanto, R., Prasetyanti, D. N., Vikasari, C., dan Pratiwi, A. F. Infotekmesin,	*Pemanfaatan Algoritma Random Forest Regression dalam Memprediksi Kepuasan Mahasiswa terhadap Dosen*	2025	Data penilaian mahasiswa; Random Forest Regression, seleksi fitur, grid search, dan cross-validation.	Random Forest Regression dapat digunakan untuk memprediksi target numerik. Pemilihan fitur, penyetelan parameter, dan validasi silang mendukung pembentukan model yang sesuai dengan data.

No	Peneliti	Judul	Tahun	Data dan Metode di gunakan	Hasil Penelitian
5	Sembiring, A., dan Santoso, H. Sinkron(Sembiring & Santoso, 2026)	*Comparative Academic Performance Prediction in Primary Schools Using Linear Regression and Random Forest*	2026	Data kinerja akademik siswa; Linear Regression dan Random Forest Regression.	Model sederhana dapat memberikan kinerja yang lebih sesuai daripada Random Forest pada struktur data tertentu. Temuan ini mendukung penggunaan model pembanding dan evaluasi objektif.

2.11. Kerangka Penelitian

Kerangka Penelitian menghubungkan kerangka KDD dengan prosedur prediksi. Data pelanggan terlebih dahulu diseleksi dan diperiksa kualitasnya. Lima prediktor dan satu target kemudian ditetapkan. Data dibagi menjadi training dan testing sebelum *one-hot encoding*. Transformasi kategorikal, penerusan DAYA_VA, dan Random Forest Regression disatukan dalam pipeline.



Gambar 2.2 Kerangka Penelitian

Pada data training, *GridSearchCV* dan *K-fold cross-validation* digunakan untuk memilih kombinasi *hyperparameter* berdasarkan RMSE. Model terbaik kemudian dilatih kembali pada seluruh data training dan dievaluasi pada data testing menggunakan MAE, MSE, RMSE, R^2, serta sMAPE. Kinerja model dibandingkan dengan baseline berbasis rata-rata.

Tahap terakhir menggunakan *permutation importance* sebanyak 20 pengulangan dengan perubahan RMSE sebagai dasar penilaian. Hasil prediksi, perbandingan dengan baseline, dan importance setiap prediktor diinterpretasikan dalam konteks data penelitian. Kesimpulan dibatasi pada kemampuan prediktif model dan tidak digunakan untuk menyatakan hubungan sebab akibat.