PROYEK AKHIR UAS

BIG DATA AND DATA MINING (ST168)

ANALISIS PREDIKSI BIAYA MEDIS PERSONAL MENGGUNAKAN ALGORITMA RANDOM FOREST REGRESSION

Disusun oleh:

23.11.5524

Ahmad Natsrul Ulum

23S1IF-BigData3

PROGRAM STUDI S1 INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

2026

1. PENDAHULUAN

1.1. Latar Belakang

Industri asuransi kesehatan menghadapi tantangan besar dalam menjaga keseimbangan antara profitabilitas perusahaan dan keterjangkauan premi bagi nasabah. Salah satu kunci utama dalam manajemen risiko asuransi modern adalah kemampuan untuk mengestimasi biaya medis (*medical costs*) di masa depan secara akurat. Penelitian terbaru menunjukkan bahwa penerapan teknologi *Machine Learning* untuk prediksi biaya kesehatan dapat membantu perusahaan mengurangi kerugian finansial akibat klaim yang tidak terduga, serta meningkatkan efisiensi operasional [1].

Biaya medis seseorang dipengaruhi oleh berbagai faktor yang kompleks, mulai dari demografi (usia, jenis kelamin) hingga gaya hidup (indeks massa tubuh, kebiasaan merokok). Metode statistik konvensional seringkali kesulitan menangkap hubungan non-linear yang rumit antara variabel-variabel tersebut. Beberapa studi komparatif membuktikan bahwa penggunaan algoritma *Forest Regression* memberikan tingkat akurasi yang lebih tinggi dibandingkan metode regresi linear tradisional dalam memprediksi biaya asuransi yang fluktuatif [2]. Selain itu, pendekatan adaptif dalam pemrosesan data juga dinilai efektif untuk menangani variasi profil risiko nasabah yang beragam [3].

Faktor gaya hidup dan ketersediaan data kesehatan terbuka (*Open Healthcare Data*) kini menjadi variabel krusial dalam analisis tren biaya pasien [4]. Oleh karena itu, diperlukan model komputasi yang mampu mempelajari pola risiko tersebut dari data historis. Algoritma *Machine Learning* menjadi pilihan populer dalam domain ini karena ketahanannya terhadap *overfitting* dan kemampuannya menangani interaksi antar fitur yang kompleks [5]. Penelitian lain juga menunjukkan bahwa *Random Forest* yang dikombinasikan atau dibandingkan dengan *Multiple Linear Regression* mampu menghasilkan prediksi premi yang lebih optimal dan *robust* [6].

1.2. Tujuan

Tujuan utama dari penelitian proyek akhir ini adalah:

1. Mengidentifikasi faktor-faktor dominan yang mempengaruhi besaran tagihan biaya medis personal.
2. Membangun model prediktif menggunakan algoritma *Random Forest Regression* untuk mengestimasi biaya medis (*charges*).
3. Mengevaluasi performa model menggunakan metrik statistik (*MAE, RMSE, R2 Score*) untuk memastikan validitas prediksi.

1.3. Metode

Penelitian ini menerapkan metodologi standar *CRISP-DM (Cross-Industry Standard Process for Data Mining)* yang terdiri dari tahapan:

1. Business Understanding: Memahami kebutuhan bisnis asuransi dalam prediksi risiko.
2. Data Understanding: Eksplorasi karakteristik dataset asuransi.
3. Data Preprocessing: Pembersihan data dan transformasi variabel kategorikal.
4. Modeling: Pelatihan model menggunakan algoritma Random Forest.
5. Evaluation: Pengukuran akurasi model menggunakan data uji.

2. PROFILE DATASET

2.1. Sumber Data

Dataset yang digunakan dalam penelitian ini adalah "**Medical Cost Personal Datasets**". Dataset ini merupakan data sekunder publik yang diperoleh dari repositori Kaggle (Link Dataset). Dataset ini dipilih karena memenuhi kriteria data publik yang valid dan belum pernah dibahas dalam studi kasus perkuliahan sebelumnya.

2.2. Karakteristik Data

Dataset ini terdiri dari 1.338 entri data (baris) dan 7 atribut (kolom) tanpa data ganda. Berikut adalah rincian atribut data:

1. **age** (Numerik): Usia penerima manfaat utama (Nasabah).
2. **sex** (Kategorikal): Jenis kelamin kontraktor asuransi (female, male).
3. **bmi** (Numerik): Indeks massa tubuh (Body Mass Index), indikator berat badan ideal (kg/m^2). Rata-rata BMI dalam data ini adalah 30.66.
4. **children** (Numerik): Jumlah anak/tanggungan yang ditanggung oleh polis.
5. **smoker** (Kategorikal): Status merokok nasabah (yes, no).
6. **region** (Kategorikal): Wilayah tempat tinggal di AS (northeast, southeast, southwest, northwest).
7. **charges** (Numerik - Target): Biaya medis individu yang ditagihkan oleh asuransi kesehatan.

3. DATA PREPROCESSING

3.1. Penanganan Missing Values

Langkah awal preprocessing adalah memeriksa kelengkapan data. Berdasarkan pemeriksaan menggunakan fungsi **isnull().sum()**, **tidak ditemukan nilai yang hilang (*missing values*)** pada seluruh kolom dataset. Oleh karena itu, data dinyatakan bersih dan siap digunakan tanpa perlu melalui proses imputasi.

3.2. Encoding Data Kategorikal

Sebagian besar algoritma *Machine Learning* matematis memerlukan input berupa angka. Oleh karena itu, fitur kategorikal diubah menjadi numerik menggunakan teknik **Label Encoding**:

1. Fitur **sex**: Diubah menjadi 0 (female) dan 1 (male).
2. Fitur **smoker**: Diubah menjadi 0 (no) dan 1 (yes).
3. Fitur **region**: Diubah menjadi angka 0, 1, 2, dan 3 mewakili empat wilayah berbeda.

3.3. Ulasan Metode

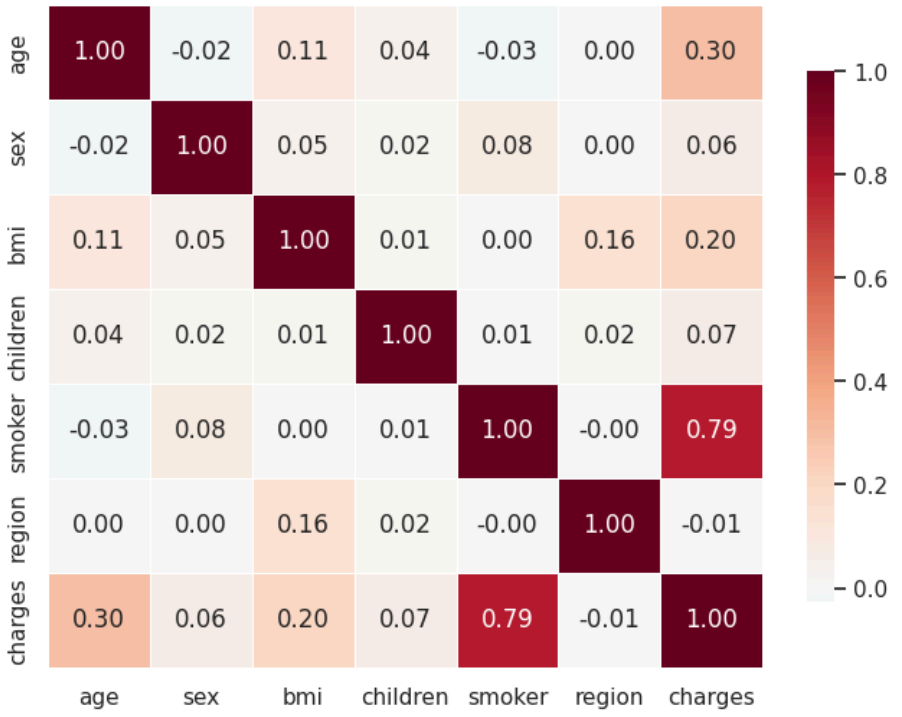
Saya memilih metode *Label Encoding* karena efisien dalam menjaga dimensi data tetap ringkas. Algoritma berbasis pohon *(Tree-based)* seperti Random Forest mampu memproses data ordinal hasil *Label Encoding* dengan sangat baik tanpa bias urutan, berbeda dengan algoritma berbasis jarak seperti KNN yang mungkin memerlukan *One-Hot Encoding*.

4. EXPLORATORY DATA ANALYSIS

4.1. Analisis Korelasi (Heatmap)

Untuk memahami hubungan antar variabel, dilakukan visualisasi matriks korelasi Pearson.

Gambar 1. Matriks Korelasi Pearson

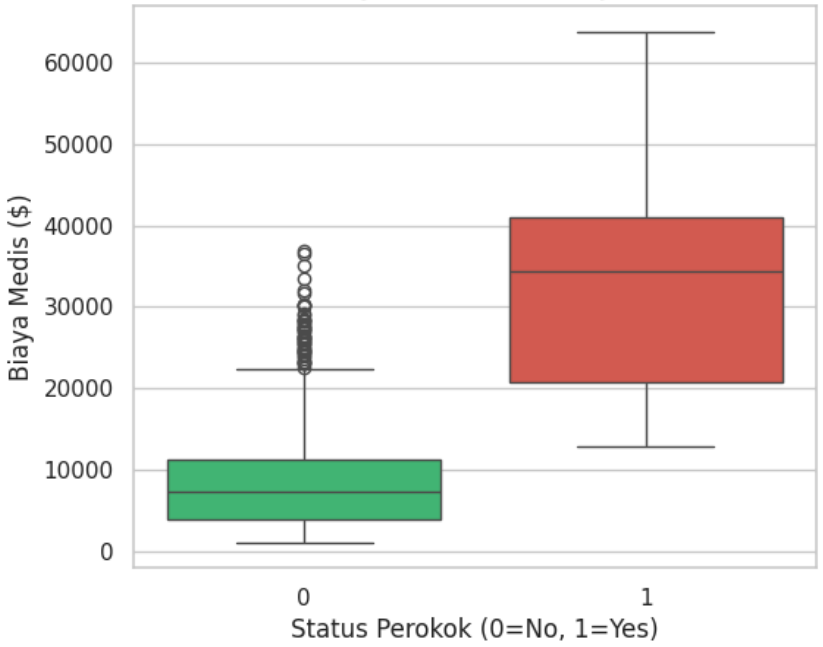


Berdasarkan heatmap di atas, terlihat bahwa fitur smoker memiliki korelasi positif yang sangat kuat (0.79) terhadap target charges. Hal ini mengindikasikan bahwa status merokok adalah faktor paling dominan yang menaikkan biaya medis. Fitur age (0.30) dan bmi (0.20) juga memiliki pengaruh positif moderat.

4.2. Distribusi Biaya (Boxplot)

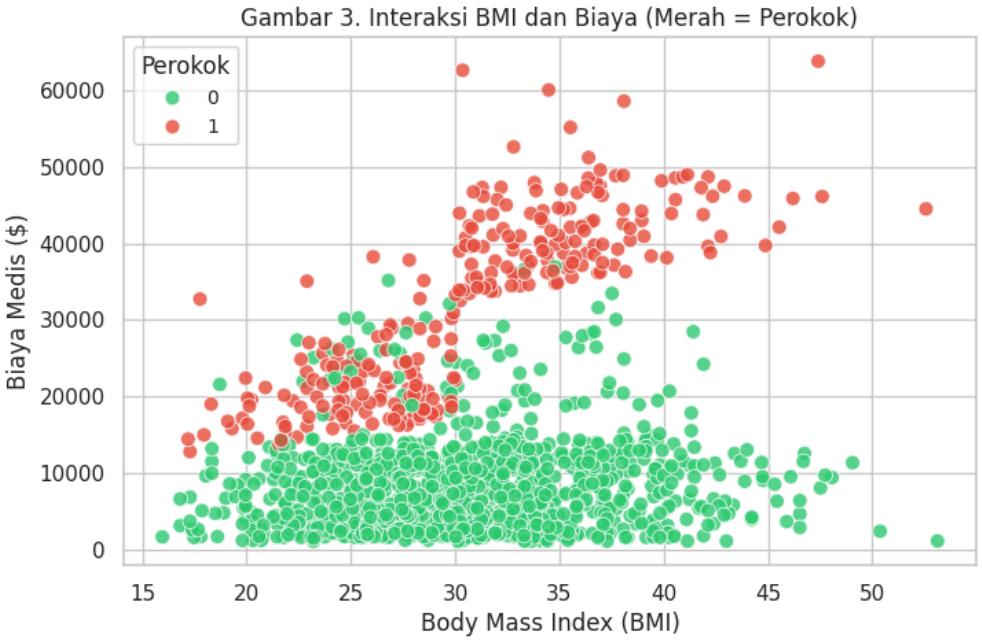
Visualisasi ini bertujuan melihat sebaran biaya antara kelompok perokok dan nonperokok.

Gambar 2. Distribusi Biaya: Non-Perokok (Hijau) vs Perokok (Merah)



Grafik Boxplot memperlihatkan perbedaan signifikan. Median biaya medis untuk perokok (kode 1) berada di kisaran di atas $30,000, sedangkan untuk non-perokok (kode 0) rata-rata hanya di bawah $10,000. Ini mengkonfirmasi bahwa perokok adalah segmen "berisiko tinggi".

4.3. Interaksi Fitur (Scatterplot)



Scatterplot di atas mengungkapkan pola interaksi yang menarik. Titik-titik berwarna merah (perokok) dengan BMI tinggi cenderung berkumpul di bagian paling atas grafik. Ini menunjukkan bahwa nasabah yang merokok DAN memiliki obesitas akan dikenakan biaya medis yang ekstrem, jauh lebih tinggi daripada sekadar penjumlahan risiko keduanya.

5. SELEKSI FITUR

5.1. Proses Seleksi

Dalam penelitian ini, fitur yang digunakan adalah seluruh atribut yang tersedia, yaitu: **age**, **sex**, **bmi**, **children**, **smoker**, dan **region**. Target variabel adalah **charges**.

5.2. Ulasan

Meskipun fitur seperti **region** dan **sex** memiliki korelasi statistik yang relatif rendah terhadap **target**, keputusan diambil untuk tetap menggunakan seluruh fitur. Alasannya, Algoritma *Random Forest* memiliki mekanisme seleksi fitur internal (*feature importance*) yang bekerja otomatis saat pembentukan pohon keputusan. Membuang fitur secara manual berisiko menghilangkan informasi interaksi yang halus (misalnya, jenis kelamin mungkin tidak berpengaruh secara umum, tetapi berpengaruh signifikan pada kelompok umur tertentu).

6. MODELING

6.1. Konfigurasi Model

Model prediktif dibangun menggunakan algoritma **Random Forest Regressor** dari pustaka Scikit-Learn dengan konfigurasi parameter sebagai berikut:

1. **Algoritma**: Random Forest (Ensemble Learning).
2. **Parameter**: n_estimators = 100 (Menggunakan 100 pohon keputusan), random_state = 42 (Untuk konsistensi hasil eksperimen).
3. **Pembagian Data**: Dataset dibagi menjadi **80% Data Latih (Train)** dan **20% Data Uji (Test)**.

6.2. Ulasan Pemilihan Model

Random Forest dipilih karena karakteristik dataset yang memuat hubungan non-linear yang kuat (seperti interaksi BMI-Perokok). Model regresi linear biasa cenderung *underfit* pada kasus seperti ini, sedangkan Random Forest mampu membentuk batas keputusan yang kompleks dan lebih akurat [7].

6.3. Tautan Proyek

1. Link Repository GitHub (Model): https://github.com/arrul37/medical-cost-prediction.git
2. Link Notebook (Google Colab): https://colab.research.google.com/drive/1fk658GBawVQ42OCZahK6yk0874PDJYRz?usp=sharing
3. Link Launchinpad: https://launchinpad.amikom.ac.id/project/analisis-prediksi-biaya-medis-personal-menggunakan-algoritma-random-forest-regression-f824193
4. Link Video Youtube: https://youtu.be/xf7moRQ1Tnc

7. EVALUASI MODEL

7.1. Hasil Evaluasi

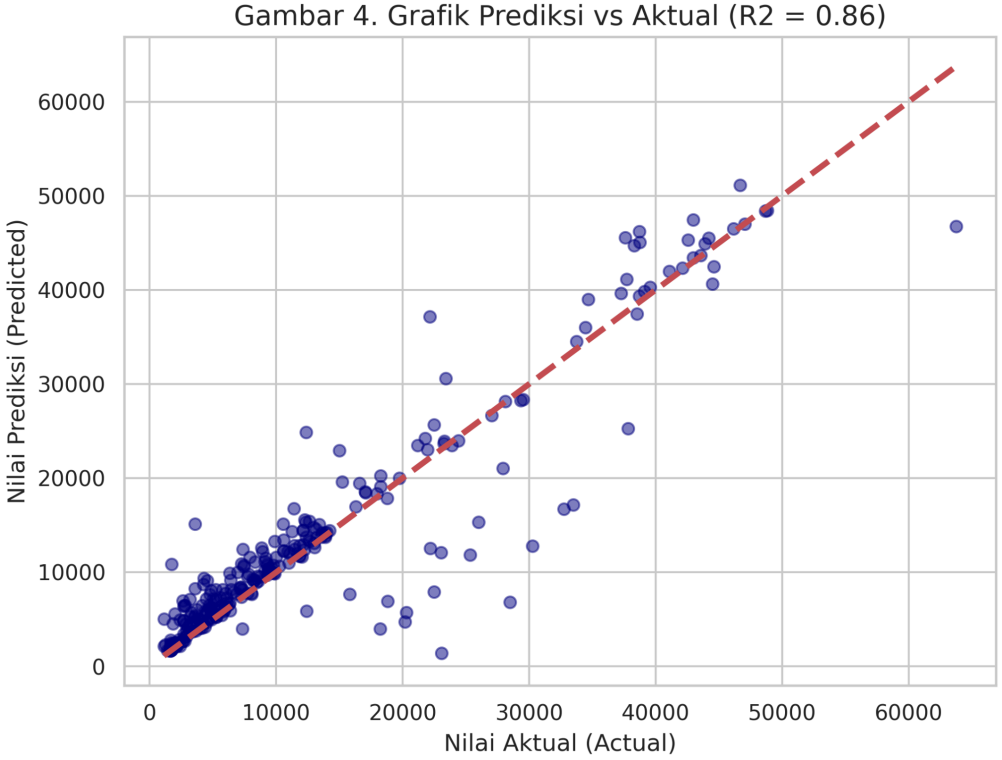
Evaluasi dilakukan terhadap data testing (20% data) yang tidak dilihat model saat pelatihan. Berikut hasilnya:

Matrik Evaluasi	Nilai	Keterangan
MAE (Mean Absolute Error)	2,533.67	Rata-rata prediksi meleset sebesar $2.533
RMSE (Root Mean Squared Error)	4,590.57	Tingkat kesalahan standar model
R2 Score	8,643	Tingkat akurasi (Good Fit)

7.2. Interpretasi

Nilai R2 Score sebesar 0.8643 menunjukkan bahwa model Random Forest mampu menjelaskan 86,43% variasi dari data biaya medis. Dalam analitika prediktif kesehatan, skor di atas 0.80 dikategorikan sebagai model yang sangat baik.

Gambar 4. Grafik Prediksi vs Aktual (R2 = 0.86)



8. ANALISA DAN PEMBAHASAN

Berdasarkan rangkaian eksperimen dan evaluasi model yang telah dilakukan, terdapat beberapa temuan analitis utama yang dapat dibahas:

8.1. Dominasi Fitur "Smoker"

Keberhasilan model mencapai akurasi R2 Score sebesar 0.86 (86%) sangat dipengaruhi oleh variabel status perokok (smoker). Analisis Feature Importance dari model Random Forest menempatkan variabel ini sebagai fitur dengan bobot tertinggi (> 0.60). Hal ini mengkonfirmasi hipotesis awal bahwa dalam penentuan premi asuransi, gaya hidup (perokok aktif) adalah faktor risiko yang jauh lebih deterministik dibandingkan faktor biologis murni seperti usia atau jenis kelamin.

8.2. Validasi Melalui Simulasi Kasus (Case Study)

Untuk memvalidasi logika model dalam skenario dunia nyata, dilakukan simulasi prediksi manual terhadap dua profil nasabah dengan atribut identik (Usia 19 tahun, Wanita, BMI 27.9), namun dengan status perokok yang berbeda. Hasil simulasi menunjukkan disparitas biaya yang ekstrem:

a. **Nasabah Perokok**: Diprediksi memiliki tagihan estimasi sebesar **±$17,000.**
b. **Nasabah Non-Perokok**: Diprediksi hanya memiliki tagihan sebesar **±$2,000.**

Selisih biaya yang mencapai hampir 8 kali lipat ini membuktikan bahwa model berhasil menangkap pola risiko kesehatan secara presisi, di mana status perokok memberikan penalti biaya yang sangat signifikan.

8.3. Keunggulan Algoritma Ensemble

Pada tahap Exploratory Data Analysis (EDA), terlihat bahwa hubungan antara BMI dan Biaya Medis bersifat non-linear (biaya hanya melonjak tajam jika nasabah memiliki BMI tinggi dan merokok). Model linear tradisional (seperti Multiple Linear Regression) cenderung gagal menangkap interaksi bersyarat seperti ini. Sebaliknya, Random Forest Regression bekerja dengan mekanisme pohon keputusan (decision trees) yang mampu membuat aturan logika percabangan ("JIKA Perokok DAN Obesitas MAKA Mahal"), sehingga error prediksi dapat diminimalisir secara signifikan dibandingkan model linear biasa.

9. KESIMPULAN

Berdasarkan rangkaian eksperimen yang telah dilakukan, dapat disimpulkan bahwa:

1. Faktor yang paling signifikan mempengaruhi tingginya biaya asuransi kesehatan adalah status merokok, diikuti oleh usia dan BMI.
2. Penerapan algoritma Random Forest Regression terbukti sangat efektif untuk kasus ini, menghasilkan performa yang tinggi dengan R2 Score 0.8643.
3. Model yang dihasilkan memiliki tingkat akurasi yang dapat diandalkan dan layak digunakan sebagai alat bantu (decision support system) bagi perusahaan asuransi untuk mengestimasi premi nasabah baru secara otomatis [8].

10. REFERENSI

[1] K. Bhatia, S. S. Gill, N. Kamboj, M. Kumar, and R. K. Bhatia, “Health Insurance Cost Prediction using Machine Learning,” in *2022 3rd International Conference for Emerging Technology (INCET)*, IEEE, May 2022, pp. 1–5. Accessed: Jan. 20, 2026. [Online]. Available: https://doi.org/10.1109/incet54531.2022.9824201

[2] V. Ramachandran, A. R. Kavitha, and R. Pandimeena, “An Accurate Prediction of Medical Insurance Cost Using Forest Regression Algorithms,” in *2023 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*, IEEE, Dec. 2023, pp. 1–4. Accessed: Jan. 20, 2026. [Online]. Available: https://doi.org/10.1109/icdsaai59313.2023.10452541

[3] S. Ahmad, A. Agarwal, and H. Ansari, “Prediction of Insurance Premium using Machine Learning with an Adaptive Approach,” in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, Jul. 2023, pp. 1–5. Accessed: Jan. 20, 2026. [Online]. Available: https://doi.org/10.1109/icccnt56998.2023.10307009

[4] A. Ravishankar Rao, R. Jain, M. Singh, and R. Garg, “Machine Learning Models For Patient Medical Cost Prediction and Trend Analysis Using Open Healthcare Data,” in *2023 IEEE 3rd International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB)*, IEEE, Apr. 2023, pp. 292–296. Accessed: Jan. 20, 2026. [Online]. Available: https://doi.org/10.1109/iceib57887.2023.10170453

[5] R. D, M. S. K, and D. J, “Health Insurance Cost Prediction using Machine Learning Algorithms,” in *2022 International Conference on Edge Computing and Applications (ICECAA)*, IEEE, Oct. 2022, pp. 1381–1384. Accessed: Jan. 20, 2026. [Online]. Available: https://doi.org/10.1109/icecaa55415.2022.9936153

[6] T. Zhang, “Prediction for Insurance Premiums Based on Random Forest and Multiple Linear Regression,” *BCP Business & Management*, vol. 38, pp. 2315–2321, Mar. 2023, doi: 10.54691/bcpbm.v38i.4097.

[7] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Elsevier, 2011.

[8] A. Pajankar and A. Joshi, “Introduction to Machine Learning with Scikit-learn,” in *Hands-on Machine Learning with Python*, Berkeley, CA: Apress, 2022, pp. 65–77. Accessed: Jan. 20, 2026. [Online]. Available: https://doi.org/10.1007/978-1-4842-7921-2_5