PROJEK AKHIR UAS

BIG DATA AND DATA MINING

PREDIKSI KEBUTUHAN KONSUMSI AIR HARIAN BERBASIS MACHINE LEARNING MENGGUNAKAN ALGORITMA RANDOM FOREST REGRESSOR

Disusun oleh

23.61.0251

Anwar Fauzi

23BCI01

PROGRAM STUDI S1 INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

2025/2026

1. PENDAHULUAN

Air merupakan komponen esensial bagi tubuh manusia yang berfungsi untuk menjaga keseimbangan cairan, mengatur suhu tubuh, dan melancarkan metabolisme. Kebutuhan air harian setiap individu sangat bervariasi dan dipengaruhi oleh faktor intrinsik seperti usia, berat badan, dan jenis kelamin, serta faktor ekstrinsik seperti tingkat aktivitas fisik dan kondisi cuaca [1]. Ketidaktepatan dalam mengonsumsi air dapat berakibat fatal; kekurangan cairan menyebabkan dehidrasi yang menurunkan performa kognitif dan fisik [2], sementara konsumsi berlebih dapat menyebabkan hiponatremia.

Metode konvensional dalam menentukan kebutuhan air seringkali hanya menggunakan aturan umum (seperti "8 gelas sehari") yang tidak terpersonalisasi. Oleh karena itu, pendekatan berbasis data (*data driven*) diperlukan untuk memberikan estimasi yang lebih presisi. Penelitian sebelumnya telah menunjukkan bahwa teknologi *Machine Learning* efektif dalam memprediksi parameter kesehatan dan kebutuhan nutrisi secara personal [3].

Tujuan dari proyek ini adalah membangun model *Predictive Analytics* menggunakan algoritma Random Forest Regressor untuk memprediksi jumlah liter air yang harus dikonsumsi seseorang per hari (*Daily Water Intake*). Metode ini dipilih karena kemampuannya dalam menangani hubungan non-linear yang kompleks antara variabel fisik dan lingkungan [4], serta ketahanannya terhadap *overfitting* dibandingkan metode regresi linier biasa. Diharapkan model ini dapat menjadi dasar bagi pengembangan sistem rekomendasi kesehatan yang cerdas [5].

2. PROFILE DATASET

Dataset yang digunakan dalam penelitian ini adalah "Daily Water Intake Dataset".

- Sumber Data: Dataset diunggah secara mandiri (Daily_Water_Intake.csv) dan merupakan data sintetis yang merepresentasikan parameter kesehatan masyarakat umum.
- Jumlah Data: Dataset terdiri dari 30.000 baris (records) dan 7 kolom (features).
- Karakteristik Atribut:

    1. Age (Numerik): Usia individu dalam tahun.
    2. Gender (Kategorikal): Jenis kelamin (Male/Female).
    3. Weight (kg) (Numerik): Berat badan individu dalam kilogram.

4. Physical Activity Level (Ordinal): Tingkat aktivitas fisik (Low, Moderate, High).
5. Weather (Ordinal): Kondisi cuaca saat aktivitas dilakukan (Cold, Normal, Hot).
6. Daily Water Intake (liters) (Numerik - Target): Jumlah konsumsi air harian dalam liter.
7. Hydration Level (Kategorikal): Status hidrasi (Good/Poor).

- Kualitas Data: Tidak ditemukan *missing value* (data kosong) pada dataset ini, sehingga data siap untuk diproses lebih lanjut.

3. DATA PREPROCESSING

Tahap *preprocessing* dilakukan untuk mempersiapkan data mentah agar dapat diterima oleh algoritma *Machine Learning*. Langkah-langkah yang dilakukan meliputi:

1. Encoding Variabel Kategorikal:
    - Label Encoding: Diterapkan pada fitur Gender (Male=0, Female=1) karena hanya memiliki dua nilai biner.
    - Ordinal Mapping: Diterapkan pada Physical Activity Level (Low=0, Moderate=1, High=2) dan Weather (Cold=0, Normal=1, Hot=2). Pendekatan ini dipilih untuk mempertahankan informasi urutan/tingkatan intensitas yang penting bagi model regresi.
2. Pemisahan Fitur dan Target: Memisahkan kolom target (Daily Water Intake) dari fitur prediktor.
3. Data Splitting: Membagi dataset menjadi 80% data latih (Train Set) untuk melatih model dan 20% data uji (Test Set) untuk evaluasi, guna memastikan pengujian yang objektif.
4. Feature Scaling: Menggunakan StandardScaler untuk menstandarisasi fitur numerik (Age, Weight) agar memiliki rata-rata 0 dan deviasi standar 1. Hal ini membantu algoritma konvergensi lebih stabil.

4. EXPLORATORY DATA ANALYSIS (EDA)

Berdasarkan analisis visualisasi data yang dilakukan:

1. Analisis Korelasi (Heatmap): Ditemukan korelasi positif yang sangat kuat antara Weight (Berat Badan) dan Daily Water Intake. Artinya, semakin berat tubuh seseorang, semakin tinggi kebutuhan airnya. Korelasi positif juga terlihat pada fitur Physical Activity Level dan Weather, di mana aktivitas tinggi dan cuaca panas meningkatkan kebutuhan air.
2. Distribusi Target: Histogram menunjukkan bahwa distribusi data Daily Water Intake berbentuk lonceng (distribusi normal). Hal ini mengindikasikan bahwa data target seimbang dan sangat cocok dimodelkan menggunakan metode Regresi, meminimalkan risiko bias pada nilai ekstrem.

5. SELEKSI FITUR

Proses seleksi fitur dilakukan berdasarkan logika domain (*domain knowledge*) dan analisis korelasi:

- Fitur yang Dibuang: Kolom Hydration Level dihapus dari fitur input.
    - *Alasan:* Variabel ini adalah "dampak" atau label hasil dari konsumsi air, bukan penyebab. Menggunakannya sebagai input akan menyebabkan *Data Leakage* (kebocoran jawaban), di mana model "mencontek" hasil akhir untuk membuat prediksi.
- Fitur Terpilih: Age, Gender, Weight (kg), Physical Activity Level, dan Weather. Kelima fitur ini dipilih karena merupakan faktor kausalitas yang secara langsung mempengaruhi kebutuhan fisiologis cairan tubuh manusia.

6. MODELING

- Metode: Algoritma yang digunakan adalah Random Forest Regressor.
- Alasan Pemilihan: Random Forest adalah metode *ensemble* yang membangun banyak pohon keputusan (*decision trees*) dan mengambil rata-rata prediksinya. Metode ini dipilih karena sangat tangguh terhadap *noise*, mampu menangkap pola hubungan non-linear yang kompleks (seperti hubungan antara cuaca dan aktivitas), dan memiliki akurasi yang umumnya lebih tinggi dibandingkan *Decision Tree* tunggal.
- Konfigurasi: Model dibangun dengan parameter n_estimators=100 (100 pohon keputusan).
- Simpan Model: Model telah disimpan dengan nama water_intake_predictor_model.pkl.
- Link Github:

    - https://github.com/Anwarfauzi03/water-intake-prediction/blob/dbdf6b812bd428eb98a25f6553e0c805104768d3/UAS_FAUZI_23_61_0251%20(1).ipynb
- Link Launchpad:
    - https://launchinpad.amikom.ac.id/project/projek-akhir-uas-big-data-and-data-mining-b279524
- Link Colab
    - https://colab.research.google.com/drive/1yzja_kubV0Hv7dZpjcEW5VzW8cUWqK_P?usp=sharing
- Link YouTube:
    -

7. EVALUASI MODEL

Evaluasi dilakukan menggunakan data uji (20% data) dengan metrik sebagai berikut:

1. R-squared (R2 Score): 0.8525
    - *Interpretasi:* Model berhasil menjelaskan sekitar 85.25% variasi data kebutuhan air. Nilai ini menunjukkan bahwa model memiliki performa yang sangat baik (*Good Fit*) dalam mengenali pola data.
2. RMSE (Root Mean Squared Error): 0.3181
    - *Interpretasi:* Rata-rata kesalahan prediksi model adalah sekitar 0.31 liter. Angka ini relatif kecil dibandingkan rentang total konsumsi air, menandakan prediksi model cukup presisi untuk penggunaan sehari-hari.

8. ANALISA DAN PEMBAHASAN

Berdasarkan hasil *Feature Importance* dari model Random Forest, ditemukan urutan prioritas faktor yang mempengaruhi prediksi:

1. Weight (Berat Badan): Menjadi faktor paling dominan. Secara biologis, ini valid karena volume air tubuh berbanding lurus dengan massa tubuh.
2. Weather (Cuaca): Faktor kedua terpenting. Suhu lingkungan secara langsung memicu mekanisme termoregulasi (keringat), yang meningkatkan kebutuhan penggantian cairan.

3. Physical Activity Level: Aktivitas fisik membakar kalori dan mengeluarkan cairan, sehingga menjadi faktor penentu ketiga.

Model berhasil memetakan bahwa kombinasi "Berat Badan Tinggi" + "Cuaca Panas" + "Aktivitas Tinggi" akan menghasilkan prediksi kebutuhan air tertinggi, sesuai dengan prinsip fisiologi kesehatan.

9. KESIMPULAN

Berdasarkan eksperimen yang dilakukan, dapat disimpulkan bahwa:

1. Algoritma Random Forest Regressor efektif digunakan untuk memprediksi kebutuhan air harian dengan akurasi 85.25% (R2 Score).
2. Faktor utama yang menentukan kebutuhan air seseorang secara berturut-turut adalah Berat Badan, Cuaca, dan Tingkat Aktivitas Fisik.
3. Model ini dapat diimplementasikan sebagai fitur pada aplikasi kesehatan pintar (*Smart Health App*) untuk memberikan rekomendasi hidrasi yang terpersonalisasi.

10. REFERENSI

[1] S. A. Kavouras and D. B. Anastasiou, "Water physiology: Essentiality, metabolism, and health implications," *Nutrition Today*, vol. 45, no. 6, pp. S27-S32, Nov. 2010.

[2] N. H. Alam *et al.*, "A Machine Learning Approach to Detect Dehydration in Afghan Children," in *arXiv preprint arXiv:2305.13275*, 2023.

[3] C. Dai, R. Chen, and C. Lin, "A fuzzy recommendation system for daily water intake," in *International Journal of Distributed Sensor Networks*, vol. 12, no. 6, pp. 1-11, Jun. 2016.

[4] J. D. Adams *et al.*, "Personalized prediction of optimal water intake in adult population by blended use of machine learning and clinical data," *Scientific Reports*, vol. 12, no. 1, Art. no. 19692, Nov. 2022.

[5] Irwan Reza Firmansyah1 , Z. K. A. Baizal1 , and Ramanti Dharayani1, "DiabeticFoodBot: Food and Water Intake Recommender System for Diabetics," in *Proceedings of the 14th International Conference on Deep Learning*, vol. 1, pp. 312-319, 2023.