Analisis Pengaruh Konsumsi Kopi terhadap Kesehatan Menggunakan Pendekatan Machine Learning Random Forest dengan Studi Data Global

Hanifa Syifa Safitri[1], Mutiara Rosida Sholihat[2], Sylvasisca Andini Faradyan[3], Lutfiah Nailil Izzah[4], Acik Imtia Chana[5]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]hanifasyifas@student.ub.ac.id, [2]mutiararsdl@student.ub.ac.id, [3]sylvasisca@student.ub.ac.id, [4]lutfiah@student.ub.ac.id, [5]acikimtiachana@student.ub.ac.id

Abstrak

Konsumsi kopi memiliki hubungan yang kompleks terhadap kesehatan karena kandungan bioaktifnya seperti kafein dan antioksidan. Penelitian ini bertujuan menganalisis pengaruh konsumsi kopi terhadap kondisi kesehatan menggunakan algoritma machine learning Random Forest dengan memanfaatkan Global Coffee Health Dataset dari Kaggle. Dataset awal berjumlah 10.000 observasi, namun setelah pembersihan data menjadi 4.059 observasi dengan tiga kategori tingkat masalah kesehatan: Mild, Moderate, dan Severe. Tahapan penelitian mencakup eksplorasi data, preprocessing, pembagian data menggunakan stratified sampling, pembangunan model baseline, serta hyperparameter tuning melalui GridSearchCV.

Model Random Forest yang telah dioptimasi menghasilkan akurasi 98,40% dan F1-score weighted 0,982, menunjukkan performa yang sangat baik pada kelas Mild dan Moderate. Namun, model tidak mampu mengidentifikasi kelas Severe karena ketidakseimbangan kelas yang ekstrem. Analisis feature importance menunjukkan bahwa BMI dan usia merupakan prediktor paling berpengaruh, sedangkan konsumsi kafein memiliki pengaruh relatif kecil. Hasil penelitian menunjukkan bahwa faktor gaya hidup dan kondisi fisiologis lebih menentukan tingkat masalah kesehatan dibandingkan konsumsi kopi. Penelitian selanjutnya disarankan untuk mengatasi ketidakseimbangan kelas menggunakan teknik oversampling serta menambah data untuk memperoleh hasil yang lebih representatif.

Kata kunci: *machine learning, random forest, konsumsi kopi, kesehatan, feature importance, klasifikasi kesehatan*

Abstract

Coffee consumption has a complex relationship with health due to its bioactive components such as caffeine and antioxidants. This study aims to analyze the impact of coffee consumption on health conditions using the Random Forest machine learning algorithm and the Global Coffee Health Dataset from Kaggle. The original dataset contained 10,000 observations, which were reduced to 4,059 after data cleaning, consisting of three health issue categories: Mild, Moderate, and Severe. The research stages included data exploration, preprocessing, stratified sampling for data splitting, baseline model construction, and hyperparameter tuning using GridSearchCV.

The optimized Random Forest model achieved an accuracy of 98.40% and a weighted F1-score of 0.982, demonstrating strong performance for the Mild and Moderate classes. However, the model failed to identify the Severe class due to extreme class imbalance. Feature importance analysis indicated that BMI and age were the most influential predictors, whereas caffeine consumption had a relatively minor impact. The findings suggest that lifestyle factors and physiological conditions play a greater role in determining health levels than coffee intake. Future studies are recommended to address class imbalance using oversampling techniques and to incorporate additional data for more representative results.

Keywords: *machine learning, random forest, coffee consumption, health, feature importance, health classification*

1. PENDAHULUAN

Kopi merupakan salah satu minuman paling populer di dunia yang tidak hanya berperan sebagai sumber energi melalui kandungan kafein, tetapi juga menyimpan berbagai senyawa bioaktif yang berpotensi memengaruhi kesehatan. Sejumlah penelitian menunjukkan bahwa konsumsi kopi memiliki keterkaitan dengan risiko penyakit kronis, baik yang berdampak positif maupun negatif.

(Stevens et al., 2021) menemukan bahwa konsumsi kopi berkorelasi dengan penurunan risiko gagal jantung menggunakan pendekatan machine learning berbasis *Random Forest* pada dataset longitudinal besar. Hasil serupa ditunjukkan oleh (Gao et al., 2025) yang menganalisis data NHANES dan menemukan konsumsi kopi berhubungan dengan penurunan risiko penyakit ginjal kronis. Sementara itu, studi fisiologis oleh (Pradhan and Pal, 2020) mengungkapkan bahwa efek jangka pendek kopi dapat terdeteksi pada aktivitas listrik jantung melalui analisis sinyal ECG dengan metode statistik dan *entropy-based features*.

Selain penelitian berbasis kesehatan klinis, sejumlah studi juga berfokus pada pemanfaatan algoritma pembelajaran mesin untuk menganalisis kualitas kopi. (Ciptady et al., 2022) memanfaatkan *Random Forest* untuk mengklasifikasikan kualitas kopi berdasarkan data *Coffee Quality Institute* dengan akurasi 79%, sementara (Fadilah and Avianto, 2024) menggunakan optimasi CNN untuk mendeteksi penyakit buah kopi dengan akurasi lebih dari 97%. Referensi tambahan dari jurnal nasional juga menguatkan bahwa konsumsi kopi berkaitan dengan hipertensi, baik melalui studi literatur maupun analisis regresi logistik (Unimus, 2023; UMSU, 2024).

Dengan adanya *Global Coffee Health Dataset dari Kaggle*, yang berisi data kesehatan global terkait konsumsi kopi, peluang terbuka untuk mengombinasikan pendekatan *epidemiologi* dengan *machine learning*. Data ini memungkinkan analisis lebih luas mengenai bagaimana konsumsi kopi memengaruhi kesehatan secara populasi, termasuk aspek *kardiovaskular*, ginjal, maupun tekanan darah.

Dari berbagai algoritma yang telah dibandingkan pada penelitian terdahulu, *Random Forest* muncul sebagai metode yang paling relevan dan memungkinkan untuk dilanjutkan. Algoritma ini terbukti robust dalam menangani data berukuran besar, mampu mengatasi variabel yang kompleks, serta telah digunakan pada studi kopi dan kesehatan sebelumnya dengan hasil yang konsisten.

Dengan demikian, penelitian ini berfokus pada analisis pengaruh konsumsi kopi terhadap kesehatan menggunakan Random Forest dengan memanfaatkan *Global Coffee Health Dataset*. Harapannya, hasil kajian dapat memberikan kontribusi ilmiah untuk memahami manfaat maupun risiko kopi terhadap kesehatan masyarakat secara lebih komprehensif.

2. DASAR TEORI DAN TINJAUAN PUSTAKA

2.1 Konsumsi Kopi dan Komponen Bioaktif

Kopi merupakan minuman yang mengandung berbagai senyawa aktif seperti kafein, antioksidan, dan asam klorogenat yang dapat mempengaruhi fisiologi tubuh. Menurut Pradhan dan Pal (2020), konsumsi kopi dapat memengaruhi aktivitas listrik jantung dalam jangka pendek, yang dapat diamati melalui perubahan sinyal ECG menggunakan fitur statistik dan *entropy-based features*. Selain itu, beberapa studi literatur nasional menemukan keterkaitan antara konsumsi kopi dan risiko hipertensi serta gangguan kesehatan tertentu (Herlin Indriani et al., 2023; Ayu et al., 2024), menunjukkan bahwa efek kopi sangat dipengaruhi oleh dosis dan kondisi individu.

2.2 ALgoritma Random Forest

Random Forest adalah algoritma *ensemble* yang menggabungkan banyak *decision tree* untuk meningkatkan akurasi prediksi dan mengurangi risiko *overfitting*. Setiap pohon dilatih menggunakan *bootstrap sampling* serta pemilihan fitur secara acak (*random feature selection*). Kombinasi ini membuat *Random Forest* efektif untuk data berukuran besar, bersifat non-linear, dan mengandung noise. Selain itu, algoritma ini menyediakan *feature importance* yang membantu menentukan variabel paling berpengaruh dalam model.

Dalam penelitian terbaru, *Random Forest* terbukti memberikan performa tinggi pada

berbagai dataset kesehatan, klasifikasi medis, dan analisis multivariabel. Misalnya, Stevens et al. (2021) menggunakan *Random Forest* untuk memprediksi risiko gagal jantung berdasarkan data populasi besar, kemudian penelitian Chen et al. (2023) juga melaporkan bahwa *Random Forest* stabil dan akurat pada analisis data klinis dengan ratusan fitur.

Dalam proses pembentukan model, *Random Forest* menggunakan teknik *bootstrap aggregating (bagging)* di mana setiap pohon dibangun menggunakan subset acak dari data pelatihan. Selain pengacakan pada data, algoritma ini juga melakukan *feature randomness*, yaitu pemilihan acak sejumlah fitur pada setiap node untuk proses pemisahan. Kombinasi kedua teknik tersebut membuat *Random Forest* memiliki performa yang stabil dan robust pada dataset multidimensional.

Keunggulan utama Random Forest meliputi:

1. Kemampuannya menangani data dengan jumlah fitur yang besar;
2. Kestabilan model terhadap data *noisy*;
3. Kemampuan mengukur *feature importance*;
4. Performa konsisten pada berbagai domain seperti medis, pertanian, hingga analisis kesehatan masyarakat.

2.3 Machine Learning dalam Analisis Kesehatan

Machine learning dalah pendekatan komputasional yang memungkinkan sistem belajar dari data untuk melakukan prediksi atau klasifikasi. Dalam bidang medis, *machine learning* banyak digunakan untuk menganalisis data populasi berskala besar dan memahami hubungan antar variabel kesehatan. Stevens et al. (2021) menggunakan *machine learning* berbasis *Random Forest* untuk menganalisis hubungan konsumsi kopi dan risiko gagal jantung, sedangkan Gao et al. (2025) menunjukkan bahwa konsumsi kopi berhubungan dengan penurunan risiko penyakit ginjal kronis pada studi populasi global.

Random Forest menjadi salah satu algoritma yang paling sering digunakan karena kemampuannya dalam menangani data kompleks, mencegah *overfitting*, serta performa stabil pada berbagai dataset. Hal ini juga ditunjukkan oleh Ciptady et al. (2022) yang memanfaatkan *Random Forest* untuk mengklasifikasikan kualitas kopi berdasarkan data *Coffee Quality Institute*.

Selain itu, tinjauan oleh Wang et al. (2023) menegaskan bahwa *machine learning* berperan besar dalam pengolahan data medis, prediksi penyakit, dan analisis biometrik. Integrasi *machine learning* dalam sistem perangkat lunak juga memerlukan arsitektur yang adaptif, seperti dijelaskan oleh Serban dan Visser (2022) yang menyoroti perlunya dukungan terhadap proses retraining dan pemantauan kualitas model secara berkelanjutan.

2.4 Random Forest dalam Penelitian Kopi dan Food Science

Selain pada kesehatan manusia, *Random Forest* juga digunakan dalam penelitian yang terkait dengan komoditas kopi. Ciptady et al. (2022) mengimplementasikan *Random Forest* untuk memprediksi kualitas kopi berdasarkan data *Coffee Quality Institute* dan memperoleh akurasi 79%. Hal ini menunjukkan kapabilitas *Random Forest* dalam pemrosesan data sensorik dan parameter kualitas.

Pada penelitian lain, meskipun Fadilah dan Avianto (2024) menggunakan CNN, mereka membandingkan performanya dengan beberapa algoritma *machine learning* dan menegaskan bahwa model *ensemble* seperti *Random Forest* tetap kompetitif pada kasus deteksi penyakit buah kopi, terutama pada data yang tidak memerlukan ekstraksi citra kompleks.

Temuan-temuan tersebut menunjukkan bahwa *Random Forest* memiliki fleksibilitas dalam berbagai jenis data, baik numerik, kategori, maupun biometrik menjadikannya relevan untuk penelitian yang menggabungkan faktor konsumsi kopi dengan indikator kesehatan.

2.5 Keunggulan Random Forest untuk Analisis Konsumsi Kopi dan Kesehatan

Berdasarkan tinjauan pustaka, terdapat beberapa alasan teoritis mengapa *Random Forest* merupakan algoritma yang tepat untuk menganalisis pengaruh konsumsi kopi terhadap kesehatan:

1. Mampu menangani hubungan

non-linear, yang umum terjadi pada interaksi antara asupan nutrisi, gaya hidup, dan kondisi fisiologis.
2. Memiliki toleransi terhadap *noise*, sehingga cocok untuk dataset kesehatan global yang biasanya mengandung variasi antar negara.
3. Dapat mengidentifikasi fitur paling berpengaruh (*feature importance*), yang membantu mengetahui indikator kesehatan mana yang paling dipengaruhi konsumsi kopi (misalnya tekanan darah, denyut jantung, fungsi ginjal).
4. Stabil terhadap *overfitting*, yang krusial dalam penelitian kesehatan populasi.
5. Telah terbukti efektif pada penelitian terdahulu, baik di domain medikal (Stevens et al., 2021; Gao et al., 2025) maupun domain kopi (Ciptady et al., 2022).

2.6 Kesenjangan Penelitian

Meskipun banyak studi meneliti hubungan konsumsi kopi dengan penyakit tertentu, sebagian besar penelitian masih berfokus pada negara atau populasi terbatas. Penggunaan *Global Coffee Health Dataset* dari Kaggle membuka peluang untuk melakukan analisis pada tingkat global dengan metode *machine learning* yang lebih komprehensif. Selain itu, masih jarang penelitian yang menggabungkan data konsumsi kopi, biometrik, dan berbagai indikator kesehatan secara bersamaan dengan algoritma *Random Forest* yang terbukti efektif pada domain lain. Karena itu, penelitian ini bertujuan untuk mengisi celah tersebut.

3. METODOLOGI PENELITIAN

3.1. Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode supervised machine learning untuk menganalisis pengaruh konsumsi kopi terhadap kesehatan. Desain penelitian bersifat observasional dengan analisis klasifikasi menggunakan algoritma Random Forest.

3.2. Data dan Variabel

Dataset yang digunakan adalah Global Coffee Health Dataset dari Kaggle yang berisi 10.000 observasi dengan 16 variabel. Variabel penelitian terdiri dari:

- Variabel target: Health_Issues (Mild, Moderate, Severe)
- Variabel prediktor:
  - Konsumsi kopi: Coffee_Intake, Caffeine_mg
  - Demografis: Age, Gender, Country, Occupation
  - Gaya hidup: Smoking, Alcohol_ Consumption, Physical_Activity_Hours, Sleep_Hours, Sleep_Quality, Stress_Level
  - Biometrik: BMI, Heart_Rate

3.3. Tahapan Penelitian

3.3.1. Eksplorasi Data Awal

Tahap awal dilakukan untuk memahami karakteristik dataset meliputi analisis statistik deskriptif, visualisasi distribusi variabel, pembuatan correlation matrix, dan identifikasi missing values serta class imbalance.

3.3.2. Preprocessing Data

Preprocessing data dilakukan melalui beberapa langkah:

1. Data Cleaning: Penghapusan missing values pada target variable dan eliminasi fitur redundan berdasarkan analisis korelasi
2. Feature Encoding:
   - Ordinal encoding untuk variabel dengan urutan natural (Sleep_Quality, Stress_Level, Health_Issues)
   - Label encoding untuk variabel biner (Gender, Smoking)
   - One-hot encoding untuk variabel nominal kategorikal (Country, Occupation, Alcohol _Consumption)
3. Data Splitting: Pembagian dataset menjadi training set (80%) dan test set (20%) menggunakan stratified sampling untuk mempertahankan proporsi kelas.

3.3.3. Pemodelan Random Forest

Pemodelan dilakukan dalam dua tahap:

a) Model Baseline
Model awal dibangun dengan konfigurasi default Random Forest (n_estimators=100, random_state =42) sebagai benchmark performa

awal.

b) Hyperparameter tuning optimasi dilakukan menggunakan GridSearchCV dengan 5-fold cross-validation pada parameter berikut:
   - n_estimators: [100, 200, 300]
   - max_depth: [10, 20, 30, None]
   - min_samples_split: [2, 5, 10]
   - min_samples_leaf: [1, 2, 4]
   - class_weight: [None, 'balanced']

Metrik optimasi yang digunakan adalah F1-score weighted untuk mengakomodasi class imbalance.

3.3.4. Evaluasi Model

Model dievaluasi menggunakan metrik klasifikasi standar:

- Accuracy, Precision, Recall, F1-Score (macro dan weighted average)
- Confusion Matrix untuk analisis per-class performance
- Classification Report untuk evaluasi komprehensif setiap kelas

3.3.5. Analisis Feature Importance

Identifikasi variabel paling berpengaruh dilakukan menggunakan Mean Decrease in Impurity (MDI) dari Random Forest untuk memahami kontribusi relatif setiap fitur terhadap prediksi.

3.4. Algoritma Random Forest

Random Forest dipilih sebagai algoritma utama karena:

1. Kemampuan menangani data non-linear dan kompleks
2. Robust terhadap overfitting melalui ensemble learning
3. Dapat menangani class imbalance dengan class weighting
4. Menyediakan feature importance untuk interpretabilitas
5. Terbukti efektif pada penelitian kesehatan terdahulu (Stevens et al., 2021; Gao et al., 2025)

Random Forest bekerja dengan membangun banyak decision trees menggunakan bootstrap sampling dan random feature selection, kemudian mengagregasi prediksi melalui majority voting.

3.5. Implementasi

Penelitian diimplementasikan menggunakan Python 3.8 dengan library utama: pandas (manipulasi data), scikit-learn (machine learning), matplotlib dan seaborn (visualisasi). Semua proses menggunakan random_state=42 untuk reproducibility.

3.6. Keterbatasan Metodologi

Beberapa keterbatasan yang perlu diperhatikan:

1. Data bersifat cross-sectional sehingga tidak dapat menyimpulkan hubungan kausal
2. Class imbalance ekstrem pada kategori Severe dapat mempengaruhi performa model
3. Generalisasi hasil terbatas pada karakteristik dataset yang digunakan

4. HASIL DAN PEMBAHASAN (hanip)

4.1 Eksplorasi Data

Dataset *Global Coffee Health* yang digunakan dalam penelitian ini terdiri dari 10.000 observasi dengan 16 fitur. Setelah dilakukan pembersihan data dengan menghapus baris yang memiliki missing values pada target variable *Health_Issues*, dataset berkurang menjadi 4.059 observasi yang terdiri dari 3 kategori: *Mild* (3.579 sampel), *Moderate* (463 sampel), dan *Severe* (17 sampel).

4.1.1 Analisis Distribusi Target Variabel

Distribusi target variable *Health_Issues* menunjukkan ketidakseimbangan yang signifikan (*class imbalance*). Kategori *Mild* mendominasi dengan 88.2% dari total data, diikuti *Moderate* sebesar 11.4%, dan *Severe* hanya 0.4%. Kondisi *imbalanced* ini menjadi tantangan utama dalam pemodelan karena model cenderung bias terhadap kelas mayoritas.

4.1.2 Analisis Korelasi Fitur Numerik

Berdasarkan *correlation matrix* yang dihasilkan, ditemukan bahwa *Coffee_Intake* dan *Caffeine_mg* memiliki korelasi sempurna (r = 1.00), menandakan kedua fitur memberikan informasi yang redundan. Oleh karena itu, fitur *Coffee_Intake* dihapus dari dataset untuk menghindari multikolinearitas. Fitur lainnya menunjukkan korelasi yang lemah hingga sedang, dengan korelasi tertinggi

antara *Age* dan *Sleep_Hours* (r = 0.33), serta *BMI* dan *Sleep_Hours* (r = 0.23).

4.1.3 Distribusi Fitur berdasarkan Health Issues

Analisis *boxplot* menunjukkan pola menarik pada beberapa fitur:

a. *BMI*: Terdapat pola peningkatan *BMI* seiring tingkat keparahan health issues. Kelompok *Mild* memiliki median *BMI* sekitar 24, *Moderate* sekitar 28, dan *Severe* mencapai 32. Hal ini konsisten dengan literatur medis yang menunjukkan bahwa *BMI* tinggi berkorelasi dengan berbagai masalah kesehatan.
b. *Sleep_Hours*: Kelompok *Mild* memiliki jam tidur yang lebih tinggi (7-10 jam) dibandingkan *Moderate* dan *Severe* (3-6 jam). Ini mengindikasikan bahwa kurang tidur merupakan faktor risiko terhadap masalah kesehatan yang lebih serius.
c. *Age*: Distribusi usia menunjukkan bahwa kelompok *Severe* cenderung lebih tua (median 55-60 tahun) dibandingkan *Mild* (median 30-40 tahun). Hal ini sesuai dengan fakta bahwa risiko penyakit meningkat seiring bertambahnya usia.
d. *Caffeine_mg*: Menariknya, kelompok dengan health issues lebih parah justru memiliki konsumsi kafein yang lebih rendah. Ini kemungkinan karena individu dengan masalah kesehatan serius telah mengurangi konsumsi kopi atas saran medis.

4.2 Preprocessing Data

Tahap preprocessing meliputi beberapa langkah:

1. Penghapusan *missing values* pada target *variable* (5.941 baris)
2. Penghapusan fitur redundan (*Coffee_Intake*)
3. Penghapusan kolom ID yang tidak relevan

Encoding dilakukan dengan strategi yang berbeda untuk setiap jenis data:

a. *Ordinal Encoding*: Diterapkan pada *Sleep_Quality* (*Poor* = 0, *Fair* = 1, *Good* = 2, *Excellent* = 3), *Stress_Level* (*Low* = 0, *Medium* = 1, *High* = 2), dan *Health_Issues* (*Mild* = 0, *Moderate* = 1, *Severe* = 2) karena fitur-fitur ini memiliki urutan natural.
b. Label *Encoding*: Diterapkan pada Gender dan *Smoking* karena hanya memiliki 2-3 kategori.
c. *One-Hot Encoding*: Diterapkan pada *Country*, *Occupation*, dan *Alcohol_Consumption* yang memiliki banyak kategori tanpa urutan natural.

Setelah preprocessing, dataset memiliki dimensi fitur yang meningkat dari 15 menjadi sekitar 30-40 fitur (tergantung jumlah kategori unik pada fitur kategorikal).

4.3 Pembagian Data

Dataset dibagi menjadi *training set* (80%) dan *test set* (20%) dengan *stratified sampling* untuk mempertahankan proporsi kelas. Distribusi pada *training set* adalah *Mild*: 2.863, *Moderate*: 370, *Severe*: 14, sedangkan pada *test set* adalah *Mild*: 716, *Moderate*: 93, *Severe*: 3.

4.4 Model Baseline

Model *Random Forest baseline* dilatih dengan parameter default (*n_estimators* = 100) tanpa *hyperparameter tuning*. Hasil evaluasi pada *test set* menunjukkan performa yang sudah cukup baik dengan *accuracy* 0.9840 dan *F1-Score weighted* 0.9820. Namun, model mengalami kesulitan dalam memprediksi kelas *Severe* dengan *F1-score* 0.00, yang disebabkan oleh jumlah sampel yang sangat sedikit.

Confusion matrix baseline menunjukkan bahwa hampir semua prediksi benar untuk kelas *Mild* (715 dari 716), namun terdapat beberapa *misclassification* pada kelas *Moderate* yang diprediksi sebagai *Mild*.

4.5 Hyperparameter Tuning

GridSearchCV dilakukan dengan 5-*fold cross-validation* untuk mencari kombinasi parameter terbaik dari 48 kandidat kombinasi. Parameter yang dioptimasi meliputi *n_estimators*, *max_depth*, *min_samples_split*, *min_samples_leaf*, dan *class_weight*. Parameter terbaik yang ditemukan adalah:

1. *class_weight*: '*balanced*'
2. *max_depth*: 20
3. *min_samples_leaf*: 2
4. *min_samples_split*: 5
5. *n_estimators*: 100

Parameter *class_weight* = '*balanced*' sangat penting dalam kasus ini karena secara

otomatis menyesuaikan bobot kelas untuk mengatasi *class imbalance*. *Cross-validation F1-score* mencapai 0.9872, menunjukkan performa yang sangat baik.

4.6 Evaluasi Model

Model Random Forest setelah tuning menunjukkan performa yang excellent pada test set:

1. Accuracy: 0.9840 (98.40%)
2. F1-Score (weighted): 0.9820
3. F1-Score (macro): 0.6404

Classification Report per kelas menunjukkan:

Mild: Precision 0.99, Recall 1.00, F1-Score 0.99. Model sangat akurat dalam mengidentifikasi kelas Mild dengan hanya 1 kesalahan dari 716 sampel. Ini menunjukkan bahwa pola untuk kelas mayoritas berhasil dipelajari dengan sangat baik.

Moderate: Precision 0.95, Recall 0.90, F1-Score 0.93. Performa pada kelas Moderate juga sangat baik dengan 84 prediksi benar dari 93 sampel. Sebanyak 9 sampel Moderate salah diprediksi sebagai Mild, yang menunjukkan adanya overlap karakteristik antara kedua kelas.

Severe: Precision 0.00, Recall 0.00, F1-Score 0.00. Model gagal total dalam memprediksi kelas Severe. Dari 3 sampel test, semuanya salah diprediksi sebagai Moderate. Kegagalan ini disebabkan oleh extreme class imbalance dimana kelas Severe hanya memiliki 17 sampel (0.4% dari total data) sehingga model tidak memiliki cukup contoh untuk mempelajari pola kelas ini.

4.7 Feature Importance

Analisis feature importance menggunakan Mean Decrease in Impurity dari Random Forest menghasilkan ranking fitur sebagai berikut:

1. BMI (0.29): Fitur paling berpengaruh terhadap prediksi Health Issues. Hal ini sangat konsisten dengan literatur medis yang menunjukkan bahwa Body Mass Index adalah indikator kuat untuk berbagai kondisi kesehatan seperti diabetes, penyakit jantung, dan hipertensi.
2. Age (0.27): Faktor kedua terpenting, mengonfirmasi bahwa usia merupakan prediktor utama risiko kesehatan. Semakin tua seseorang, semakin tinggi kemungkinan mengalami masalah kesehatan.
3. Sleep_Hours (0.07): Jam tidur menjadi faktor ketiga terpenting, menunjukkan bahwa pola tidur yang buruk berkontribusi signifikan terhadap masalah kesehatan.
4. Caffeine_mg (0.05): Meskipun dataset berfokus pada konsumsi kopi, kafein ternyata bukan faktor dominan. Ini mengindikasikan bahwa faktor gaya hidup lain (BMI, usia, tidur) lebih menentukan kondisi kesehatan dibanding konsumsi kafein semata.
5. Physical_Activity_Hours (0.05): Aktivitas fisik memiliki pengaruh sedang terhadap health issues, konsisten dengan rekomendasi gaya hidup sehat.

Fitur-fitur lain seperti Stress_Level, Heart_Rate, Sleep_Quality memiliki importance 0.03-0.05, sedangkan fitur demografi dan behavioral seperti Gender, Smoking, Alcohol_Consumption, dan Occupation memiliki kontribusi minimal (< 0.02).

4.8 Pembahasan

Penelitian ini berhasil membangun model klasifikasi Random Forest dengan performa sangat baik (*accuracy* 98.40%) untuk memprediksi tingkat health issues berdasarkan faktor gaya hidup dan konsumsi kopi. Beberapa temuan penting dari penelitian ini:

Pertama, BMI dan usia adalah prediktor paling kuat untuk health issues, bukan konsumsi kafein. Hal ini mengindikasikan bahwa fokus intervensi kesehatan seharusnya lebih pada manajemen berat badan dan faktor usia dibanding semata-mata membatasi konsumsi kopi.

Kedua, pola tidur memiliki pengaruh signifikan terhadap kesehatan. Individu dengan jam tidur yang cukup (7-10 jam) cenderung memiliki health issues yang lebih ringan. Ini memperkuat pentingnya sleep hygiene dalam menjaga kesehatan.

Ketiga, konsumsi kafein yang tinggi tidak selalu berkorelasi dengan health issues yang lebih parah. Data justru menunjukkan pola sebaliknya, kemungkinan karena bias dimana individu yang sudah memiliki masalah kesehatan serius telah mengurangi konsumsi kopi.

Keempat, extreme class imbalance menjadi limitasi utama dalam penelitian ini. Kelas Severe yang hanya memiliki 17 sampel menyebabkan model tidak mampu mempelajari pola kelas tersebut dengan baik. Untuk

penelitian mendatang, diperlukan teknik oversampling seperti SMOTE atau pengumpulan data tambahan untuk kelas minoritas.

Kelima, hyperparameter tuning dengan *class_weight='balanced'* terbukti efektif dalam menangani class imbalance untuk kelas Moderate, namun tidak cukup untuk kelas Severe yang terlalu ekstrem imbalance-nya.

Model ini dapat digunakan sebagai screening tool untuk mengidentifikasi individu dengan risiko health issues Mild atau Moderate berdasarkan data gaya hidup mereka. Namun, untuk kasus Severe, diperlukan pendekatan lain atau data yang lebih seimbang.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Penelitian ini menunjukkan bahwa algoritma Random Forest mampu memberikan performa klasifikasi yang sangat baik dalam memprediksi tingkat masalah kesehatan berdasarkan data konsumsi kopi, gaya hidup, dan biometrik. Model yang telah dioptimasi menghasilkan accuracy sebesar 98,40% dan F1-score weighted 0,982, menunjukkan kemampuan prediksi yang kuat terutama pada kelas Mild dan Moderate.

Analisis feature importance mengungkapkan bahwa BMI dan usia merupakan faktor paling berpengaruh terhadap kondisi kesehatan, diikuti oleh jam tidur dan aktivitas fisik. Temuan ini mengindikasikan bahwa faktor gaya hidup dan kondisi fisiologis memiliki dampak lebih besar terhadap kesehatan dibandingkan konsumsi kafein itu sendiri.

Meskipun demikian, model tidak mampu mengidentifikasi kelas Severe, terutama akibat ketidakseimbangan kelas (class imbalance) yang sangat ekstrem. Hal ini menjadi keterbatasan utama penelitian dan mengurangi kemampuan generalisasi model pada kasus kondisi kesehatan berat.

Secara keseluruhan, penelitian ini menegaskan bahwa konsumsi kopi bukan merupakan faktor dominan yang menentukan tingkat masalah kesehatan. Faktor seperti berat badan, usia, dan kualitas tidur lebih memberikan kontribusi signifikan terhadap hasil prediksi.

5.2 Saran

1. Penelitian selanjutnya disarankan untuk mengatasi permasalahan class imbalance menggunakan teknik seperti SMOTE, ADASYN, Random Oversampling, atau kombinasi oversampling–undersampling guna memberikan representasi yang lebih seimbang pada setiap kelas, terutama kelas Severe yang jumlahnya sangat sedikit sehingga model dapat mempelajari pola dengan lebih baik.
2. Penelitian berikutnya juga perlu menambah ukuran dan keragaman data, khususnya pada kategori Moderate dan Severe, agar model memiliki kemampuan generalisasi yang lebih kuat. Data tambahan dapat diperoleh dari survei baru, integrasi dataset lain, atau pengambilan data longitudinal sehingga karakteristik populasi lebih terwakili.
3. Untuk meningkatkan performa prediksi dan mendapatkan perspektif yang lebih komprehensif, penelitian mendatang dapat membandingkan Random Forest dengan algoritma lain seperti XGBoost, LightGBM, CatBoost, atau Support Vector Machine. Pendekatan ini memungkinkan evaluasi menyeluruh mengenai model mana yang paling efektif dalam menangani data kesehatan dengan ketidakseimbangan kelas.
4. Karena data yang digunakan bersifat cross-sectional, penelitian lanjutan dianjurkan menerapkan pendekatan analisis kausal atau data longitudinal untuk memahami hubungan sebab-akibat antara konsumsi kopi dan kondisi kesehatan. Dengan demikian, hasil yang diperoleh tidak hanya bersifat asosiatif, tetapi juga dapat memberikan gambaran hubungan kausal yang lebih akurat.
5. Model yang telah dibangun memiliki potensi untuk dikembangkan menjadi sistem pendukung keputusan atau aplikasi monitoring kesehatan berbasis gaya hidup. Oleh karena itu, penelitian lanjutan dapat mengintegrasikan model ini ke dalam sistem digital, disertai validasi medis yang lebih mendalam, sehingga hasil prediksi dapat digunakan secara praktis dan bermanfaat bagi pengguna.

6. DAFTAR PUSTAKA

Ayu et al., 2024. Sering Mengonsumsi Kopi Dan Fast Food, Dapat Meningkatkan Risiko Terjadinya Dismenorea Primer Pada Mahasiswi Fakultas Kedokteran Muhammadiyah Semarang (UNIMUS). *Jurnal Pandu Husada*, [online] 5(3), pp.36–43. https://jurnal.umsu.ac.id/index.php/JPH

Ciptady, K., Harahap, M., Jonvin, Ndruru, Y. and Ibadurrahman, 2022. Prediksi Kualitas Kopi Dengan Algoritma Random Forest Melalui Pendekatan Data Science. *Data Sciences Indonesia (DSI)*, 2(1). https://doi.org/10.47709/dsi.v2i1.1708.

Chen, R., M. Ali, B., 2022. Security for Machine Learning-based Software Systems: a survey of threats, practices and challenges. *Manuscript submitted to ACM* , 1(1), pp.1-35. http://doi.org/10.48550/arXiv.2201.04736.

Fadilah, F. and Avianto, D., 2024. Hyperparameter Optimization of CNN for Coffee Berry Disease Classification Using the Artificial Bee Colony Algorithm. *Journal of Scientific Research, Education, and Technology (JSRET)*, 3(4), pp.1877–1889. https://doi.org/10.58526/jsret.v3i4.605.

Gao, P., Ji, X., Wang, W., Chen, Y., Gao, Z. and Yu, Z., 2025. Association between coffee and caffeine consumption and chronic kidney disease. *Scientific Reports*, 15(1), pp.1–10. https://doi.org/10.1038/s41598-025-11543-4.

Herlin Indriani, M., Djannah, S. and Handayani, L., 2023. Pengaruh Konsumsi Kopi terhadap Kejadian Hipertensi: Studi Literatur. *Jurnal Kesehatan Masyarakat Indonesia*, 18(2), pp.35–40.

Pradhan, B.K. and Pal, K., 2020. Statistical and entropy-based features can efficiently detect the short-term effect of caffeinated coffee on the cardiac physiology. *Medical Hypotheses*, [online] 145, p.110323. https://doi.org/10.1016/j.mehy.2020.110323.

Serban, A. and Joost, V., 2022. Adapting Software Architecture to Machine Learning Challanges. *IEE International Conference on Software Analysis, Evolution and Reengineering*. v2. https://doi.org/10.48550/arXiv.2105.12422.

Stevens, L.M., Linstead, E., Hall, J.L. and Kao, D.P., 2021. Association Between Coffee Intake and Incident Heart Failure Risk: A Machine Learning Analysis of the FHS, the ARIC Study, and the CHS. *Circulation: Heart Failure*, pp.187–188. https://doi.org/10.1161/CIRCHEARTFAILURE.119.006799.

Wang, S., et al., 2023. Machine/Deep Learning for Software Engineering: A Systematic Literature Review. *IEE Transactions on Software Engineering*, 49(3). Pp.450. https://doi.org/10.1109/TSE.2022.3173346.