

UJIAN AKHIR SEMESTER

DATA MINING LANJUT



Dosen Pengampuh :

Muhammad Ali Akbar, M.Kom

Disusun Oleh :

- | | |
|-------------------------------------|-------------|
| 1. Dwi NurCahyo | 41522010087 |
| 2. Muhammad Farhan
Ardiansyah | 41522010264 |
| 3. Muhammad Daffa
Aulia Ramadhan | 41522010246 |
| 4. Muhammad Satrio
Dewantoro | 41522010117 |
| 5. Fawwaz Sholehuddin | 41522010239 |

Link Github : <https://github.com/Dwinurcahyo2/Data-Mining-Lanjut.git>

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS MERCU BUANA

JAKARTA

2025

BAB I PENDAHULUAN

1.1 Latar Belakang

Elektrokardiogram (EKG) merupakan salah satu instrumen diagnostik paling fundamental dan non-invasif dalam dunia kedokteran untuk memantau kesehatan jantung. Dengan merekam aktivitas kelistrikan jantung, EKG memungkinkan para ahli medis untuk mendeteksi berbagai kelainan, terutama aritmia atau gangguan irama jantung yang berpotensi fatal. Keandalan dan akurasi diagnosis yang didasarkan pada sinyal EKG menjadi pilar utama dalam menentukan tindakan medis selanjutnya, sehingga kualitas sinyal yang terekam sangatlah krusial.

Namun, dalam praktiknya, sinyal EKG sangat rentan terhadap kontaminasi dari berbagai sumber gangguan yang dikenal sebagai derau (noise) dan artefak. Gangguan ini dapat berasal dari pergerakan otot pasien, interferensi listrik dari peralatan sekitar, atau pergeseran posisi elektroda. Kehadiran derau ini dapat secara signifikan merusak integritas sinyal, menutupi fitur-fitur penting gelombang EKG, atau bahkan menghasilkan pola yang keliru menyerupai anomali jantung. Akibatnya, interpretasi yang salah dapat terjadi, yang berisiko pada kesalahan diagnosis dan penanganan pasien.

Menyadari tantangan tersebut, penelitian ini bertujuan untuk mengatasi masalah fundamental tersebut dengan mengidentifikasi keberadaan derau pada sinyal EKG. Untuk mencapai tujuan ini, analisis akan dilakukan dengan memanfaatkan dua dataset standar industri: *MIT-BIH Arrhythmia Database* yang digunakan sebagai referensi sinyal bersih, dan *MIT-BIH Noise Stress Test Database* sebagai representasi sinyal yang telah terkontaminasi oleh berbagai jenis derau. Dengan membandingkan kedua dataset ini, sebuah model dapat dilatih untuk mengenali pola derau secara efektif.

Kemampuan untuk mengidentifikasi dan memisahkan derau dari sinyal EKG merupakan langkah krusial, tidak hanya untuk meningkatkan akurasi klinis tetapi juga untuk inovasi teknologi kesehatan di masa depan. Secara jangka panjang, model yang andal dari penelitian ini diharapkan dapat diimplementasikan pada perangkat keras (*hardware*) portabel. Hal ini akan membuka jalan bagi pengembangan alat EKG yang tangguh dan terjangkau, yang dapat dioperasikan secara efektif bahkan di lingkungan terbatas seperti daerah pedalaman, di mana akses terhadap peralatan canggih dan tenaga ahli medis seringkali minim.

1.2 Rumusan Masalah

Berdasarkan latar belakang dan metodologi yang diuraikan, penelitian ini dirancang untuk menjawab serangkaian pertanyaan spesifik yang berfokus pada pengembangan sistem identifikasi derau yang praktis dan terukur:

1. Bagaimana kombinasi teknik pra-pemrosesan sinyal, yang mencakup filtering (*Notch*, *High-pass*, *Low-pass*), dengan ekstraksi fitur statistik (seperti energi, variansi, dan *skewness*) dapat menghasilkan set data yang mampu membedakan secara efektif antara sinyal EKG bersih dan yang terkontaminasi derau?
2. Seberapa efektif model klasifikasi *RandomForestClassifier*, dengan penerapan strategi pembobotan kelas (*class_weight='balanced'*) untuk mengatasi potensi ketidakseimbangan data, dalam mempelajari pola dari fitur yang telah diekstraksi untuk mengklasifikasikan sinyal secara akurat?
3. Seberapa tinggi tingkat akurasi, presisi, dan terutama recall dari model yang telah dilatih dalam mengidentifikasi kelas 'noise' pada data uji, sebagai tolok ukur utama kesiapannya untuk mencegah analisis sinyal berkualitas buruk dalam aplikasi di dunia nyata?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah ditetapkan, tujuan dari penelitian ini diuraikan secara spesifik sebagai berikut:

1. Menerapkan serangkaian teknik pra-pemrosesan sinyal yang terdiri dari *filtering* (*Notch*, *High-pass*, dan *Low-pass*) dan dilanjutkan dengan mengekstrak serangkaian fitur statistik untuk menghasilkan sebuah dataset yang representatif dan siap digunakan untuk pemodelan.
2. Membangun dan melatih sebuah model klasifikasi menggunakan algoritma *RandomForestClassifier*, dengan mengoptimalkan penanganan data yang tidak seimbang melalui penggunaan parameter *class_weight='balanced'*.
3. Mengevaluasi secara kuantitatif performa model yang telah dilatih menggunakan metrik standar klasifikasi (akurasi, presisi, dan F1-score), dengan memberikan penekanan khusus pada pencapaian nilai *recall* yang maksimal untuk kelas 'noise'.

4. Mengembangkan sebuah alur kerja prediksi yang utuh dan dapat diaplikasikan untuk mengklasifikasikan data sinyal EKG baru yang belum pernah dilihat sebelumnya sebagai 'bersih' atau 'berderau'.

1.4 Hipotesa

Berdasarkan tujuan dan metodologi penelitian yang telah dirancang, hipotesis yang diajukan dalam analisis ini adalah sebagai berikut:

1. Efektivitas Ekstraksi Fitur untuk Diferensiasi Kelas:
Diasumsikan bahwa penerapan *filtering* (*Notch*, *High-pass*, *Low-pass*) yang dilanjutkan dengan ekstraksi fitur statistik (seperti energi, variansi, dan *zero-crossing rate*) akan berhasil mengubah sinyal EKG menjadi sebuah representasi fitur yang dapat dibedakan secara kuantitatif. Sinyal yang terkontaminasi derau dihipotesiskan akan menunjukkan nilai-nilai fitur yang secara statistik signifikan berbeda dari sinyal bersih, sehingga memungkinkan pemisahan kedua kelas tersebut.
2. Keunggulan Model dalam Klasifikasi dengan Penanganan Ketidakseimbangan:
Dihipotesiskan bahwa model *RandomForestClassifier*, dengan kemampuannya menangkap hubungan non-linier, akan mampu mempelajari batas keputusan yang kompleks dari fitur-fitur yang telah diekstrak. Lebih lanjut, penggunaan parameter `class_weight='balanced'` secara spesifik diasumsikan akan berhasil memitigasi bias terhadap kelas mayoritas ('arrhythmia'), yang akan menghasilkan performa klasifikasi yang unggul, terutama tercermin pada nilai recall yang tinggi untuk kelas minoritas ('noise').

BAB II EXPLORATORY DATA ANALYSIS (EDA)

2.1 Persiapan Data

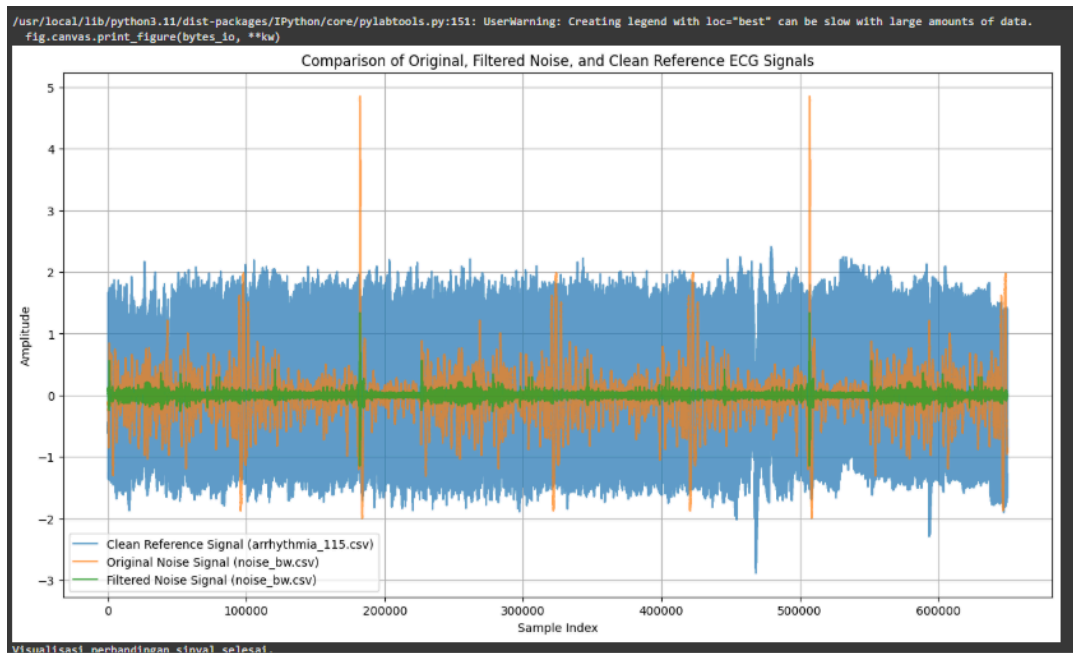
Langkah pertama dalam alur kerja ini adalah mempersiapkan data agar dapat diakses dan diproses dengan mudah. Data mentah dari kedua dataset, yaitu *MIT-BIH Arrhythmia Database* dan *MIT-BIH Noise Stress Test Database*, yang aslinya berformat non-standar (DAT/HEA), dikonversi menjadi format Comma-Separated Values (CSV). Setiap rekaman sinyal disimpan sebagai file CSV terpisah untuk memfasilitasi pemrosesan individual. File yang dihasilkan diberi nama sesuai dengan dataset asalnya (misalnya, *arrhythmia_100.csv* atau *noise_118e00.csv*) untuk kemudahan identifikasi kelas pada tahap selanjutnya.

2.2 Pra-pemrosesan Sinyal (Filtering)

Sinyal EKG yang terekam sering kali terkontaminasi oleh berbagai jenis derau yang dapat mengganggu analisis. Untuk memitigasi masalah ini, serangkaian filter digital diterapkan pada setiap sinyal dalam file CSV. Tujuan dari tahap ini adalah untuk membersihkan sinyal dari artefak yang paling umum. Filter yang digunakan adalah:

- Notch Filter (50 Hz): Diterapkan untuk secara spesifik menghilangkan interferensi dari saluran listrik (*Power Line Interference* / PLI).
- High-pass Filter (0.5 Hz): Digunakan untuk mengurangi atau menghilangkan pergeseran garis dasar (*Baseline Wander* / BW) yang sering disebabkan oleh pernapasan atau pergerakan pasien.
- Low-pass Filter (40 Hz): Berfungsi untuk meredam derau frekuensi tinggi, seperti yang berasal dari kontraksi otot (*Muscle Noise* / MN).

Hasil dari proses filtering ini disimpan sebagai file CSV baru di direktori terpisah yang kemudian menjadi input untuk tahap ekstraksi fitur.



Gambar 2.1 Visualisasi Hasil Filtering Sinyal

Menampilkan perbandingan sinyal ECG antara sinyal noise asli (oranye), sinyal setelah difilter (hijau), dan sinyal referensi bersih (biru). Terlihat bahwa proses filtering berhasil mereduksi komponen noise sehingga bentuk sinyal lebih mendekati referensi bersih. Hal ini menunjukkan keberhasilan Notch Filter, High-pass Filter, dan Low-pass Filter dalam membersihkan sinyal sebelum masuk ke tahap ekstraksi fitur.

2.3 Ekstraksi Fitur

Untuk memungkinkan model klasifikasi mempelajari pola yang membedakan antara sinyal bersih dan berderau, fitur-fitur kuantitatif diekstraksi dari setiap sinyal EKG yang telah difilter. Fitur-fitur ini berfungsi untuk meringkas karakteristik sinyal ke dalam bentuk numerik yang ringkas. Fitur yang diekstraksi meliputi:

- Energi: Mengukur kekuatan total sinyal.
- Variansi: Mengindikasikan sebaran amplitudo sinyal.
- Puncak Spektrum: Frekuensi dengan magnitudo tertinggi dalam domain frekuensi.
- Skewness: Mengukur ketidaksimetrisan distribusi amplitudo sinyal.
- Kurtosis: Mengukur keruncingan distribusi amplitudo sinyal.
- Zero-Crossing Rate: Mengukur seberapa sering sinyal berubah tanda, yang dapat mengindikasikan adanya derau frekuensi tinggi.

Setiap fitur yang dihitung dari setiap sinyal dikumpulkan ke dalam satu DataFrame tunggal. Sebuah kolom label ('arrhythmia' atau 'noise') ditambahkan ke setiap baris berdasarkan nama file sumbernya.

2.4 Pembagian Data

Langkah terakhir dalam persiapan data adalah membagi DataFrame fitur menjadi set data pelatihan dan pengujian. Pembagian ini krusial untuk melatih model dan mengevaluasi performanya secara objektif pada data yang belum pernah dilihat sebelumnya. Proses pembagian dilakukan sebagai berikut:

- Rasio Pembagian: Dataset dibagi dengan rasio 80% untuk data pelatihan dan 20% untuk data pengujian.
- Stratifikasi: Pembagian dilakukan secara stratifikasi berdasarkan label untuk memastikan bahwa proporsi kelas 'arrhythmia' dan 'noise' tetap sama di kedua set data. Hal ini penting untuk mencegah bias, terutama karena adanya ketidakseimbangan jumlah sampel antar kelas.
- Encoding Label: Label kategorikal ('arrhythmia' dan 'noise') dikonversi menjadi format numerik (0 dan 1) menggunakan LabelEncoder, karena model machine learning memerlukan input numerik.

2.5 Filtering data

Proses filtering data dilakukan dengan beberapa tujuan utama yang krusial untuk menjamin validitas hasil penelitian. Pertama, filtering berfungsi untuk meningkatkan relevansi dengan memastikan bahwa data yang digunakan sepenuhnya sesuai dengan lingkup penelitian, misalnya dengan hanya memilih data dari periode waktu tertentu, wilayah geografis spesifik, atau kategori produk yang relevan. Selanjutnya, proses ini esensial untuk meningkatkan kualitas data melalui penghapusan entri yang tidak lengkap (*missing values*), data duplikat yang dapat menyebabkan penghitungan ganda, serta data yang jelas-jelas salah (*error*). Terakhir, filtering juga bertujuan untuk memfokuskan analisis dengan cara menghapus *outlier* atau data ekstrem yang berpotensi membiaskan hasil statistik, sehingga analisis yang dilakukan dapat memberikan gambaran yang lebih akurat dan representatif.

BAB III MODEL

Bab ini menguraikan tahapan pemodelan yang dilakukan untuk mengklasifikasikan sinyal EKG ke dalam dua kelas, yaitu *arrhythmia* dan *noise*. Proses pemodelan mencakup pembuatan dataset fitur, penanganan ketidakseimbangan data, pelatihan model menggunakan algoritma Random Forest, serta optimasi melalui hyperparameter tuning untuk memperoleh performa terbaik.

3.1 Pembentukan Dataset Fitur

Setelah tahap pra-pemrosesan dan filtering sinyal, enam fitur penting berbasis domain waktu dan frekuensi diekstraksi dari masing-masing segmen sinyal EKG. Fitur-fitur yang digunakan antara lain:

- Energi
- Variansi
- Skewness
- Kurtosi
- Zero-Crossing Rate
- Puncak Spektrum

Fitur-fitur ini dihitung untuk setiap file sinyal dan digabungkan ke dalam sebuah *DataFrame* tabular, dilengkapi dengan label target ('arrhythmia' atau 'noise') berdasarkan nama file. Dataset ini kemudian di-*encode* menggunakan LabelEncoder dan dibagi menjadi data pelatihan dan pengujian dengan rasio 80:20 menggunakan stratified split untuk menjaga proporsi kelas.

3.2 Penanganan Ketidakseimbangan Data

Distribusi awal kelas menunjukkan bahwa jumlah data *arrhythmia* jauh lebih banyak dibandingkan *noise*, yang dapat menyebabkan bias model. Untuk mengatasi hal ini, digunakan *Synthetic Minority Over-sampling Technique (SMOTE)* yang mensintesis data minoritas agar jumlah sampelnya seimbang.


```

Jumlah sampel sebelum SMOTE (X_train): 100
Distribusi kelas sebelum SMOTE:
Kelas 'arrhythmia' (0): 76 sampel
Kelas 'noise' (1): 24 sampel

Jumlah sampel setelah SMOTE (X_train_res): 152
Distribusi kelas setelah SMOTE:
Kelas 'arrhythmia' (0): 76 sampel
Kelas 'noise' (1): 76 sampel

```

Gambar 3.1 Distribusi data sebelum dan sesudah SMOTE

3.3 Pelatihan Model Random Forest

Model utama yang digunakan adalah Random Forest Classifier, yaitu algoritma ensemble berbasis pohon keputusan. Model dilatih menggunakan dataset hasil SMOTE dengan parameter default untuk baseline performance.

```

Melatih model RandomForestClassifier dengan data yang sudah di-SMOTE menggunakan SEMUA FITUR...
Model RandomForestClassifier berhasil dilatih dengan data yang sudah di-SMOTE menggunakan SEMUA FITUR.

```

Gambar 3.2 Pelatihan Model Random Forest

3.4 Evaluasi Awal Model

Evaluasi awal dilakukan menggunakan data uji asli. Berikut hasil performanya:

```

Mengevaluasi performa model yang dilatih dengan data SMOTE pada data uji asli...
Prediksi pada data uji selesai.

Accuracy (after SMOTE): 0.6538

Classification Report (after SMOTE):

```

	precision	recall	f1-score	support
arrhythmia	0.82	0.70	0.76	20
noise	0.33	0.50	0.40	6
accuracy			0.65	26
macro avg	0.58	0.60	0.58	26
weighted avg	0.71	0.65	0.67	26

Gambar 3.3 Evaluasi Awal Model

3.5 Hyperparameter Tuning

Untuk meningkatkan performa, dilakukan optimasi parameter menggunakan GridSearchCV dengan skema Stratified K-Fold (n=5). Parameter yang diuji meliputi:

- n_estimators: [100, 200, 300]
- max_depth: [None, 10, 20]
- min_samples_split: [2, 5]
- min_samples_leaf: [1, 2]
- class_weight: ['balanced', None]

```
Memulai proses Hyperparameter Tuning dengan Cross-Validation menggunakan SEMUA FITUR...
Fitting 5 folds for each of 72 candidates, totalling 360 fits

Proses Hyperparameter Tuning selesai menggunakan SEMUA FITUR.
Parameter terbaik: {'class_weight': 'balanced', 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 200}
Skor F1 Weighted terbaik (dari Cross-Validation): 0.9213

Model terbaik (best_clf) sudah siap untuk dievaluasi pada data uji asli menggunakan SEMUA FITUR.
```

Gambar 3.4 Weighted terbaik

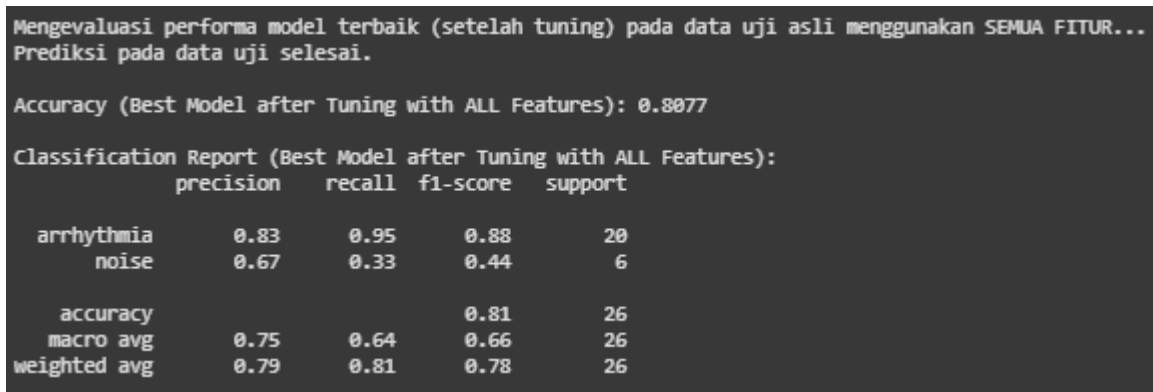
Dengan skor F1-weighted tertinggi pada data pelatihan: 0.9213.

BAB IV HASIL DAN PEMBAHASAN

Setelah dilakukan training dan tuning model Random Forest, performa dievaluasi pada data uji yang sebelumnya tidak pernah dilihat oleh model.

4.1 Hasil Evaluasi Model Terbaik

Model terbaik dari hasil tuning menunjukkan kinerja yang meningkat dibandingkan baseline awal yaitu dengan **accuracy: 0.8077**



```
Mengevaluasi performa model terbaik (setelah tuning) pada data uji asli menggunakan SEMUA FITUR...
Prediksi pada data uji selesai.

Accuracy (Best Model after Tuning with ALL Features): 0.8077

Classification Report (Best Model after Tuning with ALL Features):
      precision    recall  f1-score   support

 arrhythmia       0.83      0.95      0.88        20
      noise       0.67      0.33      0.44         6

 accuracy          0.81          0.81          0.81        26
 macro avg       0.75      0.64      0.66        26
 weighted avg    0.79      0.81      0.78        26
```

Gambar 4.1 Performa Model Terbaik

Interpretasi: Model mampu mengidentifikasi sinyal *arrhythmia* dengan sangat baik, tetapi masih menghadapi tantangan dalam mengenali sinyal *noise*. Meskipun Recall untuk *noise* rendah (0.33), nilai Precision yang tinggi (0.67) menunjukkan bahwa model cenderung berhati-hati dan hanya memprediksi *noise* saat yakin.

4.2 Dampak Tahapan Pemodelan

1. **SMOTE** berhasil menyeimbangkan kelas dan meningkatkan performa kelas minoritas (*noise*).
2. **Ekstraksi fitur yang lengkap** dari domain waktu dan frekuensi memberikan informasi yang cukup bagi model untuk membedakan dua kelas.
3. **Tuning hyperparameter** berhasil meningkatkan F1-score dan akurasi keseluruhan secara signifikan.