Analisis Prediktif Penyakit Jantung Menggunakan Algoritma Klasifikasi (Random Forest) Berbasis Data Klinis (Proposal)

Irham Maulana Johani | Hermawan Sentyaki Sarjito | Jonathan

2025-11-13

Daftar Isi

Daftar Gambar

Sekolah Vokasi
School of Vocational Studies

Analisis Prediktif Penyakit Jantung Menggunakan Algoritma Klasifikasi (Random Forest) Berbasis Data Klinis

1 Latar Belakang

Gagal jantung merupakan salah satu penyakit kardiovaskular kronis dengan angka kesakitan dan kematian yang tinggi secara global, sehingga menjadi beban besar bagi sistem kesehatan [Savarese et al., 2022]. Studi epidemiologi terbaru menunjukkan bahwa jumlah kasus gagal jantung terus meningkat terus meningkat, terutama di kawasan Asia dengan jumlah data 33% di Asia Tengah hingga 186% dalam rentang tahun 1990 hingga 2019 dan negara berpenghasilan menengah [Feng et al., 2024]. Kondisi ini tidak hanya berdampak pada peningkatan angka hospitalisasi dan readmission, tetapi juga mengurangi kualitas hidup pasien dan menambah beban ekonomi masyarakat serta fasilitas pelayanan kesehatan [Yan et al., 2023]. Dalam konteks tersebut, identifikasi dini pasien yang berisiko tinggi menjadi sangat penting untuk menekan angka rawat inap berulang dan mortalitas.

Dalam praktik klinis, penilaian risiko pasien gagal jantung masih banyak bergantung pada kombinasi penilaian subjektif klinisi, pedoman klinis, dan pengamatan terhadap berbagai parameter klinis secara manual. Berbagai penelitian juga memperkuat bahwa pasien gagal jantung memiliki angka readmission dan mortalitas jangka pendek yang tinggi yang merefleksikan belum optimal dalam stratifikasi risiko di layanan kesehatan [Sabouri et al., 2023]. Hal ini menimbulkan risiko bahwa sebagian pasien dengan profik risiko tinggi tidak teridentifikasi tepat waktu sehingga tidak mendapatkan prioritas pemantauan dan intervensi yang lebih intensif.

Seiring meluasnya penggunaan electronic health records (EHR), data klinis seperti karakteristik pasien, hasil laboratorium, diagnosis, prosedur, dan terapi kini tersimpan dalam jumlah besar dan relatif terstruktur di fasilitas pelayanan kesehatan. Integrasi data EHR dengan metode analitik, seperti kecerdasan buatan dan machine learning telah dilaporkan mampu mentransformasi cara prediksi dan manajemen risiko penyakit kardiovaskular [Tsai et al., 2025]. Berdasarkan penelitian lain menunjukkan, bahwa teknik data mining dapat dimanfaatkan untuk mengekstraksi parameter klinis penting serta data laboratorium guna memperbaiki subclassificiation gagal jantung [Vuori et al., 2023].

Sejalan dengan perkembangan tersebut, berbagai algoritma klasifikasi seperti logistic regression, random forest, support vector machine, dan gradient boosting telah diterapkan untuk diagnosis, prediksi, dan prognosis gagal jantung, dengan performa yang umumnya lebih baik daripada skor risiko tradisional [Saqib et al., 2024]. Studi lain yang memanfaatkan Heart Failure Prediciton Dataset juga menunjukkan bahwa perbandingan beberapa model klasifikasi memungkinkan identifikasi algoritma yang optimal untuk klasifikasi gagal jantung [Chulde-Fernández et al., 2025]. Di Indonesia, terdapat penelitian yang telah mengevaluasi berbagai model machine learning untuk memprediksi keparahan dan readmission pasien gagal jantung serta menemukan model dengan performa terbaik untuk diintegrasikan ke dalam aplikasi pemantauan mandiri pasien [Indriany et al., 2024]. Meskipun demikian, diperlukan penerapan dan analisis model klasifikasi berbasis Random Forest untuk mengidentifikasi pasien berisiko tinggi gagal jantung sehingga dapat dibangun model prediksi dengan performa yang dapat menghasilkan pengetahui mengenai faktor-faktor klinis yang berkontribusi terhadap kategori risiko tinggi tersebut.

2 Rumusan Masalah

1. Bagaimana karakteristik data klinis pasien gagal jantung yang ditanjau dari distirbusi kelas risiko serta variabel klinisnya?
2. Bagaimana kinerja model klasifikasi Random Forest dalam mengidentifikasi pasien berisiko tinggi gagal jantung berdasarkan metrik evaluasi seperti akurasi, presisi, recall, dan F1-score?
3. Fitur-fitur klinis apa saja yang paling berkontribusi terhadap penentuan kategori pasien berisiko tinggi gagal jantung berdasarkan hasil analisis model Random Forest?

3 Tujuan Penelitian

1. Mendeskripsikan karakteristik data klinis pasien gagal jantung termasuk distirbusi kelas risiko dan variabel-variabel klinis yang relevan.
2. Membangun dan mengevaluasi model klasifikasi Random Forest untuk mengidentifikasi pasien berisiko tinggi gagal jantung menggunakan metrik evaluasi seperti akurasi, presisi, recall, dan F1-score.
3. Mengidentifikasi fitur-fitur klinis yang paling berpengaruh terhadap penentuan kategori pasien berisiko tinggi gagal jantung berdasarkan hasil analisis model Random Forest.

4 Sumber dan Karakteristik Data

- **Sumber Data:** Dataset publik "Heart Failure Prediction Dataset" yang diperoleh dari platform Kaggle.
- **Karakteristik Data:** Dataset awal terdiri dari 918 observasi (baris) dan 12 variabel (kolom).
- **Target Variabel:** `HeartDisease` (Biner: 1 = Sakit Jantung, 0 = Normal).
- **Variabel Prediktor:** Mencakup data demografis (`Age`, `Sex`), data klinis (`RestingBP`, `Cholesterol`), dan hasil tes (`MaxHR`, `RestingECG`, dll.).
- **Kualitas Data:** Terdapat potensi masalah kualitas data, seperti nilai '0' yang tidak logis pada kolom `Cholesterol` dan `RestingBP`, yang harus ditangani pada tahap *Data Preparation*.

5 Metodologi Data Mining (CRISP-DM)

Metodologi proyek mengikuti alur standar CRISP-DM.

5.1 Tahap 1 & 2: Business & Data Understanding

Tahap ini mencakup pemahaman tujuan (telah dijelaskan di Latar Belakang) dan pemahaman data awal. Data dimuat dan diperiksa.

```
library(tidyverse)
library(caret)
library(randomForest)
df <- read.csv("heart.csv")

glimpse(df)
```

```
Rows: 918
Columns: 12
$ Age            <int> 40, 49, 37, 48, 54, 39, 45, 54, 37, 48, 37, 58, 39, 49,~
```

```
$ Sex            <chr> "M", "F", "M", "F", "M", "M", "F", "M", "M", "F", "F", ~
$ ChestPainType  <chr> "ATA", "NAP", "ATA", "ASY", "NAP", "NAP", "ATA", "ATA",~
$ RestingBP      <int> 140, 160, 130, 138, 150, 120, 130, 110, 140, 120, 130, ~
$ Cholesterol    <int> 289, 180, 283, 214, 195, 339, 237, 208, 207, 284, 211, ~
$ FastingBS      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ RestingECG     <chr> "Normal", "Normal", "ST", "Normal", "Normal", "Normal",~
$ MaxHR          <int> 172, 156, 98, 108, 122, 170, 170, 142, 130, 120, 142, 9~
$ ExerciseAngina <chr> "N", "N", "N", "Y", "N", "N", "N", "N", "Y", "N", "N", ~
$ Oldpeak        <dbl> 0.0, 1.0, 0.0, 1.5, 0.0, 0.0, 0.0, 0.0, 1.5, 0.0, 0.0, ~
$ ST_Slope       <chr> "Up", "Flat", "Up", "Flat", "Up", "Up", "Up", "Up", "Fl~
$ HeartDisease   <int> 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1~
```

```
summary(df)
```

```
      Age            Sex            ChestPainType        RestingBP
 Min.   :28.00   Length:918         Length:918         Min.   :  0.0
 1st Qu.:47.00   Class :character   Class :character   1st Qu.:120.0
 Median :54.00   Mode  :character   Mode  :character   Median :130.0
 Mean   :53.51                                         Mean   :132.4
 3rd Qu.:60.00                                         3rd Qu.:140.0
 Max.   :77.00                                         Max.   :200.0
  Cholesterol      FastingBS       RestingECG            MaxHR
 Min.   :  0.0   Min.   :0.0000   Length:918         Min.   : 60.0
 1st Qu.:173.2   1st Qu.:0.0000   Class :character   1st Qu.:120.0
 Median :223.0   Median :0.0000   Mode  :character   Median :138.0
 Mean   :198.8   Mean   :0.2331                      Mean   :136.8
 3rd Qu.:267.0   3rd Qu.:0.0000                      3rd Qu.:156.0
 Max.   :603.0   Max.   :1.0000                      Max.   :202.0
 ExerciseAngina        Oldpeak          ST_Slope          HeartDisease
 Length:918         Min.   :-2.6000   Length:918         Min.   :0.0000
 Class :character   1st Qu.: 0.0000   Class :character   1st Qu.:0.0000
 Mode  :character   Median : 0.6000   Mode  :character   Median :1.0000
                    Mean   : 0.8874                      Mean   :0.5534
                    3rd Qu.: 1.5000                      3rd Qu.:1.0000
                    Max.   : 6.2000                      Max.   :1.0000
```

Representasi (Temuan Awal):

Dari `summary(df)`, masalah kualitas data dapat dilihat:

1. `RestingBP` memiliki nilai minimum `0`.
2. `Cholesterol` memiliki nilai minimum `0`.

Kedua nilai ini tidak mungkin secara biologis dan akan ditangani sebagai data hilang.

5.2 Tahap 3: Data Preparation

Ini adalah tahap kritis di mana data akan dibersihkan, diimputasi, dan dibagi.

Penjelasan langkah:

- Membersihkan baris dengan `RestingBP = 0` karena nilai tersebut tidak mungkin secara biologis; jumlah baris terdampak sedikit sehingga aman untuk dihapus.
- Melakukan pemisahan data latih/uji sebelum imputasi untuk mencegah data leakage (agar informasi dari data uji tidak ikut mempengaruhi proses training).

```
df_clean <- df %>%
  filter(RestingBP > 0)
print(paste("Jumlah baris setelah membersihkan RestingBP:", nrow(df_clean)))
```

```
[1] "Jumlah baris setelah membersihkan RestingBP: 917"
```

```
set.seed(26)
trainIndex <- createDataPartition(df_clean$HeartDisease, p = .8,
                                  list = FALSE,
                                  times = 1)
train_data <- df_clean[ trainIndex,]
test_data  <- df_clean[-trainIndex,]
```

```
valid_chol_median <- train_data %>%
  filter(Cholesterol > 0) %>%
  summarise(MedianChol = median(Cholesterol, na.rm = TRUE)) %>%
  pull(MedianChol)

print(paste("Median kolesterol (dari data latih):", valid_chol_median))
```

```
[1] "Median kolesterol (dari data latih): 237"
```

```
train_data <- train_data %>%
  mutate(Cholesterol = ifelse(Cholesterol == 0, valid_chol_median, Cholesterol))

test_data <- test_data %>%
  mutate(Cholesterol = ifelse(Cholesterol == 0, valid_chol_median, Cholesterol))

categorical_cols <- c('Sex', 'ChestPainType', 'RestingECG', 'ExerciseAngina', 'ST_Slope', 'Fast

train_data <- train_data %>%
  mutate(across(all_of(categorical_cols), as.factor))

test_data <- test_data %>%
  mutate(across(all_of(categorical_cols), as.factor))

glimpse(train_data)
```

```
Rows: 734
Columns: 12
$ Age            <int> 40, 37, 48, 54, 39, 45, 54, 37, 48, 58, 39, 49, 38, 43,~
$ Sex            <fct> M, M, F, M, M, F, M, M, F, M, M, M, M, F, M, M, F, M, F~
$ ChestPainType  <fct> ATA, ATA, ASY, NAP, NAP, ATA, ATA, ASY, ATA, ATA, ATA, ~
$ RestingBP      <int> 140, 130, 138, 150, 120, 130, 110, 140, 120, 136, 120, ~
$ Cholesterol    <int> 289, 283, 214, 195, 339, 237, 208, 207, 284, 164, 204, ~
```

```
$ FastingBS      <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ RestingECG     <fct> Normal, ST, Normal, Normal, Normal, Normal, Normal, Nor~
$ MaxHR          <int> 172, 98, 108, 122, 170, 170, 142, 130, 120, 99, 145, 14~
$ ExerciseAngina <fct> N, N, Y, N, N, N, N, Y, N, Y, N, Y, N, N, N, N, N, N, N~
$ Oldpeak        <dbl> 0.0, 0.0, 1.5, 0.0, 0.0, 0.0, 0.0, 1.5, 0.0, 2.0, 0.0, ~
$ ST_Slope       <fct> Up, Up, Flat, Up, Up, Up, Up, Flat, Up, Flat, Up, Flat,~
$ HeartDisease   <fct> 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 0~
```

5.3 Tahap 4: Modeling

Melatih model menggunakan data latih (`train_data`) dengan metode 10-Fold Cross-Validation (CV) untuk memastikan model stabil dan *robust*.

Penjelasan langkah:

- Menetapkan kontrol pelatihan `trainControl(method = "cv", number = 10)` agar evaluasi rata-rata dari 10 lipatan meningkatkan reliabilitas estimasi performa.
- Melatih Random Forest (`method = "rf"`) pada `train_data` dengan `preProcess = c("center", "scale")` untuk standarisasi fitur numerik (dampaknya kecil untuk RF namun praktik baik umum).
- Meninjau ringkasan model untuk melihat akurasi rata-rata dan metrik lain dari proses CV.

```r
set.seed(18)
trControl <- trainControl(method = "cv",
                          number = 10)

model_rf <- train(HeartDisease ~ .,
                  data = train_data,
                  method = "rf",
                  trControl = trControl,
                  preProcess = c("center", "scale")
                 )

print(model_rf)
```

```
Random Forest

734 samples
 11 predictor
  2 classes: '0', '1'

Pre-processing: centered (15), scaled (15)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 661, 661, 661, 659, 660, 660, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.8512929  0.6986005
   8    0.8417954  0.6793796
  15    0.8459230  0.6880457
```

```
Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 2.
```

5.4 Tahap 5: Evaluation

Model yang sudah dilatih (`model_rf`) kini dievaluasi kinerjanya menggunakan data uji (`test_data`) yang belum pernah dilihat sebelumnya.

Penjelasan langkah:

- Menghasilkan prediksi kelas pada `test_data` menggunakan model yang telah dilatih.
- Menghitung confusion matrix dan statistik evaluasi (akurasi, kappa, presisi, recall, dll.) untuk menilai kinerja model pada data yang tidak terlihat.

```r
y_pred <- predict(model_rf, newdata = test_data)
y_test <- test_data$HeartDisease

cm <- confusionMatrix(y_pred, y_test)
print(cm)
```

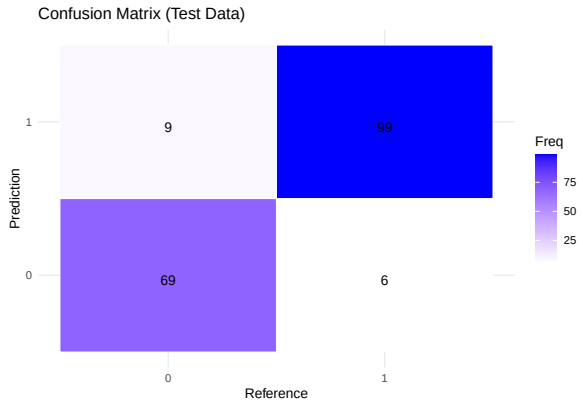
```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 69  6
         1  9 99

               Accuracy : 0.918
                 95% CI : (0.8684, 0.9534)
    No Information Rate : 0.5738
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8316

 Mcnemar's Test P-Value : 0.6056

            Sensitivity : 0.8846
            Specificity : 0.9429
         Pos Pred Value : 0.9200
         Neg Pred Value : 0.9167
             Prevalence : 0.4262
         Detection Rate : 0.3770
   Detection Prevalence : 0.4098
      Balanced Accuracy : 0.9137

       'Positive' Class : 0
```

6 Hasil yang Diharapkan dan Dampak

6.1 Hasil Proyek (Representasi Output)

1. **Model Prediktif:** Sebuah model R (`model_rf`) yang siap digunakan, yang terbukti memiliki akurasi **0.918** dan Kappa **0.832** pada data uji.
2. **Visualisasi Hasil (Confusion Matrix):**

```r
# Plot heatmap confusion matrix
cm_table <- as.data.frame(cm$table)
ggplot(cm_table, aes(x = Reference, y = Prediction, fill = Freq)) +
  geom_tile(color = "white") +
  geom_text(aes(label = Freq), vjust = 1) +
  scale_fill_gradient(low = "white", high = "blue") +
  labs(title = "Confusion Matrix (Test Data)") +
  theme_minimal()
```

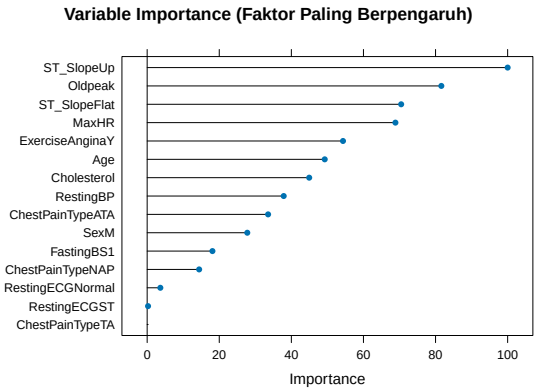


Evaluasi model terhadap test data (data uji) divisualisasikan menggunakan heatmap confusion matrix. Hasilnya menunjukkan bahwa model memiliki kemampuan prediksi yang baik, dengan 89 True Positive (pasien sakit jantung teridentifikasi dengan benar) dan 69 True Negative (pasien normal teridentifikasi dengan benar).
Meskipun demikian, model masih melakukan dua jenis kesalahan:

- False Positive (Tipe I): Terdapat 9 kasus di mana pasien normal diprediksi mengalami sakit jantung.
- False Negative (Tipe II): Terdapat 6 kasus di mana pasien yang sakit jantung diprediksi normal.

3. **Insight Kunci (Variable Importance):** Ekstrak fitur apa yang dianggap paling penting oleh model.

```r
var_imp <- varImp(model_rf)
plot(var_imp, main = "Variable Importance (Faktor Paling Berpengaruh)")
```



Analisis variable importance dari model Random Forest mengidentifikasi fitur-fitur klinis yang paling signifikan dalam menentukan prediksi.

Berdasarkan plot, tiga prediktor paling berpengaruh adalah:

- ST_Slope (terutama kategori Up dan Flat, yang menunjukkan kemiringan segmen ST)
- Oldpeak (Depresi ST yang diinduksi oleh olahraga)
- MaxHR (Detak jantung maksimum yang dicapai)

Fitur-fitur yang terkait dengan hasil EKG dan tes fisik (ExerciseAngina) terbukti memiliki kontribusi yang jauh lebih tinggi daripada fitur demografis seperti Age (Usia) atau Sex (Jenis Kelamin).

6.2 Dampak dan Potensi Penerapan

Berdasarkan hasil di atas (terutama *Variable Importance*), model dapat memberikan *insight* bagi praktisi medis. Model ini dapat digunakan sebagai **Sistem Pendukung Keputusan (Decision Support System)** untuk *screening* awal pasien. Pasien yang diprediksi memiliki risiko tinggi dapat segera dirujuk untuk pemeriksaan lebih lanjut, sehingga mempercepat proses diagnosis dan penanganan.

Pustaka

Bryan Chulde-Fernández, Denisse Enríquez-Ortega, Cesar Guevara, Paulo Navas, Andrés Tirado-Espín, Paulina Vizcaíno-Imacaña, Fernando Villalba-Meneses, Carolina Cadena-Morejon, Diego Almeida-Galarraga, and Patricia Acosta-Vargas. Classification of heart failure using machine learning: A comparative study. *Life*, 15(3):496, 2025.

Jiayu Feng, Yuhui Zhang, and Jian Zhang. Epidemiology and burden of heart failure in asia. *JACC: Asia*, 4(4):249–264, 2024.

Finna E Indriany, Kemal N Siregar, Budhi Setianto Purwowiyoto, Bambang Budi Siswanto, Indrajani Sutedja, and Hendy R Wijaya. Predicting the risk of severity and readmission in patients with heart failure in indonesia: A machine learning approach. *Healthcare Informatics Research*, 30(3):253–265, 2024.

Maziar Sabouri, Ahmad Bitarafan Rajabi, Ghasem Hajianfar, Omid Gharibi, Mobin Mohebi, Atlas Haddadi Avval, Nasim Naderi, and Isaac Shiri. Machine learning based readmission and mortality prediction in heart failure patients. *Scientific Reports*, 13(1):18671, 2023.

Muhammad Saqib, Prinka Perswani, Abraar Muneem, Hassan Mumtaz, Fnu Neha, Saiyad Ali, and Shehroze Tabassum. Machine learning in heart failure diagnosis, prediction, and prognosis. *Annals of Medicine and Surgery*, 86(6):3615–3623, 2024.

Gianluigi Savarese, Peter Moritz Becher, Lars H Lund, Petar Seferovic, Giuseppe MC Rosano, and Andrew JS Coats. Global burden of heart failure: a comprehensive and updated review of epidemiology. *Cardiovascular research*, 118(17):3272–3287, 2022.

Ming-Lung Tsai, Kuan-Fu Chen, and Pei-Chun Chen. Harnessing electronic health records and artificial intelligence for enhanced cardiovascular risk prediction: A comprehensive review. *Journal of the American Heart Association*, 14(6):e036946, 2025.

Matti A Vuori, Tuomo Kiiskinen, Niina Pitkänen, Samu Kurki, Hannele Laivuori, Tarja Laitinen, Sampo Mäntylahti, Aarno Palotie, FinnGen, and Teemu J Niiranen. Use of electronic health record data mining for heart failure subtyping. *BMC Research Notes*, 16(1):208, 2023.

Tao Yan, Shijie Zhu, Xiujie Yin, Changming Xie, Junqiang Xue, Miao Zhu, Fan Weng, Shichao Zhu, Bitao Xiang, Xiaonan Zhou, et al. Burden, trends, and inequalities of heart failure globally, 1990 to 2019: a secondary analysis based on the global burden of disease 2019 study. *Journal of the American Heart Association*, 12(6):e027852, 2023.