

**PERBANDINGAN DAN OPTIMALISASI ALGORITMA
RANDOM FOREST, C4.5, ADABOOST, DAN HIBRIDA C4.5-
ADABOOST DALAM PREDIKSI PENYAKIT HEPATITIS**



Arya Prana Sadinala

0000044294

**PROGRAM STUDI SISTEM INFORMASI
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA**

TANGERANG

2024

UNIVERSITAS
MULTIMEDIA
NUSANTARA

**PERBANDINGAN DAN OPTIMALISASI ALGORITMA
RANDOM FOREST, C4.5, ADABOOST, DAN HIBRIDA C4.5-
ADABOOST DALAM PREDIKSI PENYAKIT HEPATITIS**



SKRIPSI

Diajukan sebagai Salah Satu Syarat untuk Memperoleh
Gelar Sarjana Komputer (S.Kom)

Arya Prana Sadinala

00000044294

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS TEKNIK DAN INFORMATIKA

UNIVERSITAS MULTIMEDIA NUSANTARA

TANGERANG

2024

HALAMAN PERNYATAAN TIDAK PLAGIAT

Dengan ini saya,

Nama : Arya Prana Sadinala

Nomor Induk Mahasiswa : 00000044294

Program studi : Sistem Informasi

Skripsi dengan judul:

“PERBANDINGAN DAN OPTIMALISASI ALGORITMA RANDOM FOREST, C4.5, ADABOOST DAN HIBRIDA C4.5-ADABOOST DALAM PREDIKSI PENYAKIT HEPATITIS”

merupakan hasil karya saya sendiri bukan plagiat dari karya ilmiah yang ditulis oleh orang lain, dan semua sumber, baik yang dikutip maupun dirujuk, telah saya nyatakan dengan benar serta dicantumkan di Daftar Pustaka.

Jika di kemudian hari terbukti ditemukan kecurangan/penyimpangan, baik dalam pelaksanaan skripsi maupun dalam penulisan laporan skripsi, saya bersedia menerima konsekuensi dinyatakan TIDAK LULUS untuk Tugas Akhir yang telah saya tempuh.

Tangerang, 13 Mei 2024



Arya Prana Sadinala

U M M N
U N I V E R S I T A S
M U L T I M E D I A
N U S A N T A R A

HALAMAN PENGESAHAN

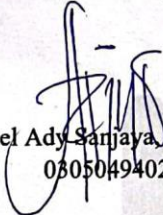
Skripsi dengan judul
**PERBANDINGAN DAN OPTIMALISASI ALGORITMA RANDOM
FOREST, C4.5, ADABOOST, DAN HIBRIDA C4.5-ADABOOST DALAM
PREDIKSI PENYAKIT HEPATITIS**

Oleh

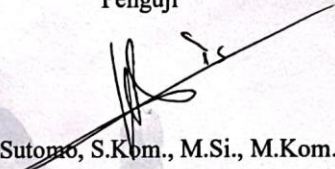
Nama : Arya Prana Sadinala
NIM : 00000044294
Program Studi : Sistem Informasi
Fakultas : Teknik dan Informatika

Telah diujikan pada hari Senin, 20 Mei 2024
Pukul 15:00 s.d 17:00 dan dinyatakan
LULUS
Dengan susunan penguji sebagai berikut.

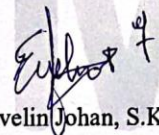
Ketua Sidang


Samuel Ady Sanjaya, S.T., M.T.
0305049402

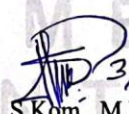
Penguji


Rudi Sutomo, S.Kom., M.Si., M.Kom.
0222057501

Pembimbing


Monika Evelin Johan, S.Kom., M.M.S.I.
0327059501/071281

Ketua Program Studi Sistem Informasi


Ririn Ikana Desanti, S.Kom., M.Kom.
313058001

LEMBAR PERSETUJUAN PUBLIKASI

KARYA ILMIAH MAHASISWA

Yang bertanda tangan dibawah ini:

Nama : Arya Prana Sadinala

Nomor Induk Mahasiswa : 00000044294

Program Studi : Sistem Informasi

Jenjang : S2 / S1 / D3

Judul Karya Ilmiah :

PERBANDINGAN DAN OPTIMALISASI ALGORITMA RANDOM FOREST, C4.5, ADABOOST, DAN HIBRIDA C4.5-ADABOOST DALAM PREDIKSI PENYAKIT HEPATITIS

Menyatakan dengan sesungguhnya bahwa saya bersedia:

Memberikan izin sepenuhnya kepada Universitas Multimedia Nusantara untuk mempublikasikan hasil karya ilmiah saya di repositori Knowledge Center, sehingga dapat diakses oleh Civitas Akademika/Publik. Saya menyatakan bahwa karya ilmiah yang saya buat tidak mengandung data yang bersifat konfidensial dan saya juga tidak akan mencabut kembali izin yang telah saya berikan dengan alasan apapun.

Saya tidak bersedia, dikarenakan:

Dalam proses pengajuan untuk diterbitkan ke jurnal/konferensi nasional/internasional (dibuktikan dengan *letter of acceptance*)*.

Tangerang, 03 Juni 2024



Arya Prana Sadinala

* Jika tidak bisa membuktikan LoA jurnal/HKI selama 6 bulan kedepan, saya bersedia mengizinkan penuh karya ilmiah saya untuk diunggah ke KC UMN dan menjadi hak institusi UMN.

KATA PENGANTAR

Puji Syukur atas selesainya penulisan Skripsi ini dengan judul: “PERBANDINGAN DAN OPTIMALISASI ALGORITMA RANDOM FOREST, C4.5, ADABOOST DAN HIBRIDA C4.5-ADABOOST DALAM PREDIKSI PENYAKIT HEPATITIS” dilakukan untuk memenuhi salah satu syarat untuk mencapainya gelar Strata1 Jurusan Sistem Informasi Pada Fakultas Teknik Informatika Universitas Multimedia Nusantara. Saya menyadari bahwa, tanpa bantuan dan bimbingan dari berbagai pihak, dari masa perkuliahan sampai pada penyusunan tugas akhir ini, sangatlah sulit bagi saya untuk menyelesaikan tugas akhir ini. Saya mengucapkan terima kasih kepada:

1. Dr. Ninok Leksono, M.A., selaku Rektor Universitas Multimedia Nusantara.
2. Dr. Eng. Niki Prastomo, S.T., M.Sc., selaku Dekan Fakultas Teknik dan Informatika Universitas Multimedia Nusantara.
3. Ririn Ikana Desanti, S.Kom., M.Kom., selaku Ketua Program Studi Sistem Informasi Universitas Multimedia Nusantara.
4. Monika Evelin Johan, S.Kom., M.M.S.I., sebagai Pembimbing yang telah memberikan bimbingan, arahan, dan motivasi atas terselesainya tugas akhir ini.
5. Johan Setiawan, S.Kom., MM., M.B.A., sebagai Dosen matakuliah MRSI yang telah memberikan bimbingan, arahan, dan motivasi atas terselesainya tugas akhir ini.
6. Keluarga saya yang telah memberikan bantuan dukungan material dan moral, sehingga penulis dapat menyelesaikan tugas akhir ini.

Semoga karya ilmiah ini dapat bermanfaat, baik dari sumber informasi maupun inspirasi bagi para pembaca.

Tangerang, 15 Mei 2024


Arya Prana Sadinala

Perbandingan dan Optimalisasi Algoritma Random Forest, C4.5,

AdaBoost dan Hibrida C4.5-AdaBoost dalam Prediksi

Penyakit Hepatitis

Arya Prana Sadinala

ABSTRAK

Hepatitis adalah kondisi peradangan hati yang disebabkan oleh berbagai faktor seperti virus, alkohol, atau penyakit autoimun. Ada beberapa jenis virus hepatitis, seperti: hepatitis a (hav), hepatitis b (hbv), hepatitis c (hcv), hepatitis d (hdv), dan hepatitis e (hev). Penting untuk dapat mengidentifikasi dan memprediksi risiko hepatitis secara dini agar tindakan pencegahan dan penanganan yang tepat dapat dilakukan. Penelitian ini bertujuan untuk memprediksi penyakit Hepatitis dengan menggunakan pendekatan *Machine Learning* berdasarkan data pasien yang diperoleh dari RSUD Kota Tangerang. Dataset yang digunakan dalam penelitian ini bersumber dari RSUD secara langsung pada Maret 2024. Dataset digunakan dengan rentang waktu bulan Januari tahun 2019 hingga bulan Desember tahun 2023 yang meliputi berbagai parameter klinis yang relevan seperti hasil rekam medis cek lab darah dan hati. Penggunaan algoritma diuji mencakup *Random Forest*, *C4.5*, *AdaBoost*, dan hibrida *C4.5-AdaBoost*. *C4.5* adalah algoritma konstruksi pohon keputusan pada bidang data mining yang sedang berkembang, tetapi telah digunakan sejak lama di berbagai bidang. *AdaBoost* adalah pendekatan pembelajaran mesin yang meningkatkan akurasi aturan prediktif dengan menggabungkan banyak aturan yang relatif lemah dan tidak akurat. Hasil penelitian menunjukkan bahwa penggunaan algoritma pemodelan tanpa optimalisasi peneliti menyarankan menggunakan Random Forest berdasarkan tingkat ROC yang sempurna, dan keempat pemodelan dengan menggunakan optimalisasi dapat menjadi pilihan lainnya, dikarenakan setiap pemodelan memiliki tingkat yang sempurna dalam memprediksi penyakit hepatitis.

Kata kunci: AdaBoost, C4.5, Penyakit Hepatitis, Hibrida C4.5-AdaBoost, Random Forest

U N I V E R S I T A S
M U L T I M E D I A
N U S A N T A R A

***Comparison and Optimization of Random Forest, C4.5, AdaBoost
and Hybrid C4.5-AdaBoost Algorithms in Hepatitis Disease***

Prediction

Arya Prana Sadinala

ABSTRACT (English)

Hepatitis is an inflammatory condition of the liver caused by various factors such as viruses, alcohol, or autoimmune diseases. There are several types of viral hepatitis, such as: hepatitis a (hav), hepatitis b (hbv), hepatitis c (hcv), hepatitis d (hdv), and hepatitis e (hev). It is important to be able to identify and predict the risk of hepatitis early so that appropriate prevention and treatment measures can be taken. This study aims to predict Hepatitis disease using Machine Learning approach based on patient data obtained from Tangerang City Hospital. The dataset used in this study was sourced from the RSUD directly in March 2024. The dataset is used with a time span of January 2019 to December 2023 which includes various relevant clinical parameters such as medical record results of blood and liver lab checks. Algorithms tested included Random Forest, C4.5, AdaBoost, and C4.5-AdaBoost hybrid. C4.5 is a decision tree construction algorithm in the emerging field of data mining but has been used for a long time in various fields. AdaBoost is a machine learning approach that improves the accuracy of predictive rules by combining many relatively weak and inaccurate rules. The results showed that the use of modeling algorithms without optimization suggested using Random Forest based on the perfect ROC level, and the four modeling using optimization can be another option, because each model has a perfect level in predicting hepatitis disease.

Keywords: *AdaBoost, C4.5, Hepatitis Disease, Hibrida C4.5-AdaBoost, Random Forest*

U N I V E R S I T A S
M U L T I M E D I A
N U S A N T A R A

DAFTAR ISI

| | |
|--|------|
| HALAMAN PERNYATAAN TIDAK PLAGIAT | ii |
| LEMBAR PERSETUJUAN PUBLIKASI | iv |
| KARYA ILMIAH MAHASISWA | iv |
| KATA PENGANTAR | v |
| ABSTRAK | vi |
| ABSTRACT (English) | vii |
| DAFTAR ISI | viii |
| DAFTAR TABEL | xi |
| DAFTAR RUMUS | xii |
| DAFTAR GAMBAR | xiii |
| DAFTAR LAMPIRAN | xv |
| BAB I PENDAHULUAN | 1 |
| 1.1 Latar Belakang | 1 |
| 1.2 Rumusan Masalah | 6 |
| 1.3 Batasan Masalah | 7 |
| 1.4 Tujuan dan Manfaat Penelitian | 7 |
| 1.4.1 Tujuan Penelitian | 7 |
| 1.4.2 Manfaat Penelitian | 7 |
| 1.5 Sistematika Penulisan | 8 |
| BAB II LANDASAN TEORI | 9 |
| 2.1 Penelitian Terdahulu | 9 |
| 2.2 Tinjauan Teori | 15 |
| 2.2.1 Hepatitis | 15 |
| 2.2.2 Rumah Sakit Umum Daerah Kota Tangerang (RSUD) | 16 |
| 2.3 Framework dan Algoritma | 16 |
| 2.3.1 Knowledge Discovery in Database (KDD) | 16 |
| 2.3.2 Algoritma Random Forest | 18 |
| 2.3.3 Algoritma C4.5 | 18 |
| 2.3.4 AdaBoost | 20 |

| | | |
|---|---|----|
| 2.3.5 | <i>KNNImputer</i> | 21 |
| 2.3.6 | <i>Hyperparameter</i> | 21 |
| 2.3.7 | <i>Confusion Matrix</i> | 21 |
| 2.3.7.1 | Akurasi..... | 22 |
| 2.3.7.2 | Presisi..... | 23 |
| 2.3.7.3 | Recall..... | 23 |
| 2.3.7.4 | F1-Score..... | 24 |
| 2.4 | Tools..... | 24 |
| 2.4.1 | <i>Google Colaboratory / Google Colab</i> | 24 |
| 2.4.2 | Python..... | 24 |
| BAB III METODOLOGI PENELITIAN | | 26 |
| 3.1 | Gambaran Umum Objek Penelitian..... | 26 |
| 3.2 | Alur Penelitian..... | 27 |
| 3.2.1 | Pre-KDD..... | 27 |
| 3.2.2 | Data Selection..... | 28 |
| 3.2.3 | Pre-Processing / Cleaning..... | 29 |
| 3.2.4 | Transformation..... | 29 |
| 3.2.5 | Data Mining..... | 30 |
| 3.2.6 | Interpratation / Evaluation..... | 30 |
| 3.3 | Metode Penelitian..... | 30 |
| 3.4 | Teknik Pengumpulan Data..... | 32 |
| 3.4.1 | Data Collection..... | 32 |
| 3.5 | Teknik Analisis Data..... | 33 |
| BAB IV ANALISIS DAN HASIL PENELITIAN | | 37 |
| 4.1 | Pre-KDD..... | 37 |
| 4.1.1 | Analisa Grafik..... | 39 |
| 4.2 | Data Selection..... | 41 |
| 4.3 | Pre-Processing..... | 42 |
| 4.4 | Tranformation..... | 45 |
| 4.5 | Data Mining..... | 48 |
| 4.5.1 | Algoritma Random Forest..... | 48 |
| 4.5.2 | Algoritma C4.5..... | 51 |

| | |
|--|----|
| 4.5.3 Algoritma AdaBoost | 54 |
| 4.5.4 Hibrida C4.5-AdaBoost | 57 |
| 4.6 Interpretation / Evaluation | 61 |
| 4.6.1 Visualisasi Perbandingan Akurasi | 61 |
| 4.6.2 Visualisasi Confusion Matrix..... | 66 |
| 4.6.3 Visualisasi Perbandingan ROC | 71 |
| 4.7 Analisis Hasil dan Pembahasan | 75 |
| BAB V SIMPULAN DAN SARAN | 77 |
| 5.1 Simpulan..... | 77 |
| 5.2 Saran..... | 78 |
| DAFTAR PUSTAKA | 80 |
| LAMPIRAN | 84 |

UMMN

UNIVERSITAS
MULTIMEDIA
NUSANTARA

DAFTAR TABEL

| | |
|--|----|
| Tabel 2. 1 Penelitian Terdahulu | 9 |
| Tabel 3. 1 Variabel Dataset..... | 28 |
| Tabel 3. 2 Perbandingan Algoritma | 31 |
| Tabel 3. 3 Perbandingan Framework | 33 |
| Tabel 3. 4 Perbandingan Bahasa Pemrograman..... | 34 |
| Tabel 3. 5 Perbandingan Tools | 35 |
| Tabel 4. 1 Tabel Perbandingan Perfoma Metrix | 75 |

v



UMMN

UNIVERSITAS
MULTIMEDIA
NUSANTARA

DAFTAR RUMUS

| | |
|--------------------------------|----|
| Rumus 2. 1 Rumus Entropy | 19 |
| Rumus 2. 2 Rumus Gain | 19 |
| Rumus 2. 3 Rumus Akurasi | 22 |
| Rumus 2. 4 Rumus Presisi | 23 |
| Rumus 2. 5 Rumus Recall..... | 23 |



UMMN

UNIVERSITAS
MULTIMEDIA
NUSANTARA

DAFTAR GAMBAR

| | |
|---|----|
| Gambar 1. 1 Grafik Kasus Kematian Hepatitis Akut Tahun 2019 | 2 |
| Gambar 1. 2 Grafik Polulasi Hepatitis B dan C di Pasifik Barat | 3 |
| Gambar 1. 3 Grafik Kematian Hepatitis di ASEAN | 4 |
| Gambar 2. 1 Knowledge Discovery in Database | 17 |
| Gambar 2. 2 Confusion Matrix | 22 |
| Gambar 3. 1 Alur Penelitian..... | 27 |
| Gambar 3. 2 Raw Dataset | 33 |
| Gambar 4. 1 Package Import..... | 37 |
| Gambar 4. 2 Rekap cek lab darah dan hati pada setiap pasien | 38 |
| Gambar 4. 3 Describe Dataset..... | 38 |
| Gambar 4. 4 Grafik Distribusi Jenis Kelamin | 39 |
| Gambar 4. 5 Grafik Distribusi Umur | 40 |
| Gambar 4. 6 Import Dataset pada Google Colab | 41 |
| Gambar 4. 7 Jumlah Missing Value di setiap Atribut..... | 42 |
| Gambar 4. 8 Skrip untuk menghilangkan missing value | 43 |
| Gambar 4. 9 Penghapusan Missing Value dengan KNNImputer | 44 |
| Gambar 4. 10 Pengecekan Kriteria setelah Handling Missing Value..... | 44 |
| Gambar 4. 11 Pembuatan Target Class | 45 |
| Gambar 4. 12 Pemilihan atribut berdasarkan uji korelasi untuk split data | 46 |
| Gambar 4. 13 Pemilihan fitur, target class, dan peningkatan data dengan Smote | 47 |
| Gambar 4. 14 Skrip split data training dan testing | 48 |
| Gambar 4. 15 Pemodelan Random Forest tanpa Hyperparameter | 49 |
| Gambar 4. 16 Pemodelan Random Forest dengan Optimalisasi..... | 50 |
| Gambar 4. 17 Hyperparameter Terbaik menurut GridSearchCV | 51 |
| Gambar 4. 18 Pemodelan C4.5 tanpa Optimalisasi..... | 52 |
| Gambar 4. 19 Pemodelan C4.5 dengan Optimalisasi..... | 53 |
| Gambar 4. 20 Hyperparameter Terbaik menurut GridSearchCV | 54 |
| Gambar 4. 21 Pemodelan AdaBoost tanpa Optimalisasi | 55 |
| Gambar 4. 22 Pemodelan AdaBoost dengan Optimalisasi | 56 |
| Gambar 4. 23 Hyperparameter Terbaik menurut GridSearchCV | 57 |
| Gambar 4. 24 Pemodelan Hibrida C4.5-AdaBoost tanpa Optimalisasi | 58 |
| Gambar 4. 25 Pemodelan Hibrida C4.5-AdaBoost dengan Optimalisasi | 59 |
| Gambar 4. 26 Hyperparameter Terbaik menurut GridSearchCV | 60 |
| Gambar 4. 27 Skrip Visualisasi Metriks tanpa Optimalisasi | 61 |
| Gambar 4. 28 Visualisasi Perbandingan Metriks tanpa Optimalisasi | 62 |
| Gambar 4. 29 Skrip Visualisasi Metriks dengan Optimalisasi | 64 |
| Gambar 4. 30 Visualisasi Perbandingan Metrix dengan Optimalisasi..... | 65 |
| Gambar 4. 31 Confusion Matrix Random Forest tanpa Optimalisasi | 66 |
| Gambar 4. 32 Confusion Matrix C4.5 tanpa Optimalisasi..... | 67 |
| Gambar 4. 33 Confusion Matrix AdaBoost tanpa Optimalisasi | 68 |
| Gambar 4. 34 Confusion Matrix hibrida C4.5-Adaboost tanpa Optimalisasi..... | 68 |

| | |
|--|----|
| Gambar 4. 35 Confusion Matrix Random Forest menggunakan Optimalisasi | 69 |
| Gambar 4. 36 Confusion Matrix C4.5 menggunakan Optimalisasi | 70 |
| Gambar 4. 37 Confusion Matrix AdaBoost menggunakan Optimalisasi..... | 70 |
| Gambar 4. 38 Confusion Matrix Hibrida C4.5-AdaBoost menggunakan Optimalisasi | 71 |
| Gambar 4. 39 Skrip Visualisasi ROC pada Pemodelan tanpa Optimalisasi | 71 |
| Gambar 4. 40 Visualisasi ROC tanpa Optimalisasi | 72 |
| Gambar 4. 41 Skrip Visualisasi ROC pada Pemodelan menggunakan Optimalisasi | 73 |
| Gambar 4. 42 Visualisasi ROC menggunakan Optimalisasi | 74 |
| Gambar 4. 43 Pengecekan Over atau Under atau Good pada Dataset setelah validasi | 76 |



DAFTAR LAMPIRAN

| | |
|---|----|
| Lampiran A Form Bimbingan Skripsi | 84 |
| Lampiran B Form Jadwal Pengambilan Data | 85 |
| Lampiran C Nametag Rekam Medis untuk Mahasiswa..... | 86 |
| Lampiran D Turnitin | 87 |



UMMN

UNIVERSITAS
MULTIMEDIA
NUSANTARA