Proposal Proyek Machine Learning: Klasifikasi Tahapan Penyakit Hati (HCV Data)

Disusun Oleh:
Gema Syahdan Prasetyo (241552010021)
Abdull Hayyi El Naseer (241552010010)

Dosen Pengampu:

Hendri Kharisma S.Kom, M.T

Program Studi Teknik Informatika
Sekolah Tinggi Manajemen Informatika dan Komputer TAZKIA
Jl.Raya Dramaga Blok Radar Baru No. 8, RT.03/RW.03, Margajaya,
Kec. Bogor Barat, Kota Bogor, Jawa Barat 16116, Indonesia

BAB 1: PENDAHULUAN

1.1 Latar Belakang Masalah

Hepatitis C Virus (HCV) merupakan salah satu penyebab utama penyakit hati kronis yang dapat berkembang menjadi fibrosis dan sirosis apabila tidak ditangani secara dini. Analisis data laboratorium dan faktor demografi pasien dapat memberikan wawasan penting dalam mendeteksi tahap perkembangan penyakit hati.

Dataset HCV Data (UCI 571) berisi data klinis 615 pasien dengan berbagai kondisi, mulai dari donor darah sehat hingga pasien dengan hepatitis, fibrosis, dan sirosis. Data ini

mencakup 12 atribut, seperti usia, jenis kelamin, serta parameter laboratorium (misalnya ALB, ALP, AST, bilirubin, CHOL, GGT, PROT, dan lainnya).

Dengan memanfaatkan teknik Supervised Machine Learning, khususnya algoritma klasifikasi multikelas, kita dapat membangun model yang mampu memprediksi tahap penyakit hati berdasarkan hasil laboratorium pasien. Model ini dapat menjadi dasar pengembangan sistem pendukung keputusan medis berbasis data.

1.2 Rumusan Masalah

1. Bagaimana cara melakukan pra-pemrosesan (preprocessing) data klinis HCV yang memiliki missing values dan variabel kategorikal agar siap untuk digunakan dalam model machine learning?

2. Seberapa akurat model Logistic Regression dan Random Forest dalam mengklasifikasikan kondisi pasien (Donor, Hepatitis, Fibrosis, Sirosis)?

3. Algoritma manakah yang menunjukkan kinerja klasifikasi terbaik berdasarkan metrik akurasi, precision, recall, dan F1-score?

1.3 Tujuan Proyek

- Melakukan eksplorasi dan pembersihan data (data cleaning & preprocessing).

- Menerapkan dua model Supervised Learning: Logistic Regression (baseline linear) dan Random Forest (ensemble non-linear).

- Mengevaluasi dan membandingkan kinerja kedua model dalam memprediksi tahap penyakit hati.

- Mengidentifikasi fitur laboratorium yang paling berpengaruh terhadap hasil klasifikasi.

BAB 2: DATASET DAN METODOLOGI

2.1 Deskripsi Dataset

Atribut	**Detail**
Nama Dataset	HCV Data
Sumber	UCI Machine Learning Repository
Link	https://archive.ics.uci.edu/dataset/571/hcv+data
Jumlah Data	615 instansi (pasien)

Jumlah Fitur	12 (usia, jenis kelamin, ALB, ALP, AST, bilirubin, CHOL, GGT, PROT, dll.)
Target Variabel (Y)	Category = {Blood Donor, Hepatitis, Fibrosis, Cirrhosis}
Tipe Data	Tabular, Clinical, Multiclass Classification
Reverensi Paper	Hybrid model for precise hepatitis-C classification using improved random forest and SVM method https://pmc.ncbi.nlm.nih.gov/articles/PMC10394001/

2.2 Metodologi Penerapan Algoritma

Kasus ini termasuk Klasifikasi Multikelas, sehingga metrik evaluasi akan difokuskan pada kemampuan model dalam memprediksi kategori pasien secara benar.

Metode	Tipe Model	Keterangan
Metode 1	Logistic Regression	Model linear baseline untuk menganalisis hubungan antara fitur laboratorium dan kategori penyakit. Cocok untuk data yang relatif terdistribusi normal.
Metode 2	Random Forest	Model ensemble berbasis pohon keputusan yang mampu menangani data non-linear dan interaksi fitur yang kompleks. Memberikan estimasi feature importance yang berguna untuk interpretasi klinis.

2.3 Rencana Preprocessing dan Feature Engineering

Langkah-langkah yang direncanakan:

1. **Eksplorasi Data:** Analisis distribusi nilai, outlier, dan proporsi tiap kategori target.

2. **Penanganan Missing Values:** Menggunakan imputasi (mean atau median) pada fitur numerik.

3. **Encoding Variabel Kategorikal:** Mengubah variabel jenis kelamin menjadi nilai numerik (mis. M = 1, F = 0).

4. **Normalisasi/Standarisasi:** Menormalkan fitur numerik agar memiliki skala yang sebanding.

5. **Split Dataset:** 70% data untuk training dan 30% untuk testing.

6. **Evaluasi:** Menggunakan metrik Accuracy, Precision, Recall, dan F1-Score. Visualisasi Confusion Matrix untuk melihat performa setiap kelas.

BAB 3: RENCANA KERJA

No	Fase Proyek	Kegiatan Utama	Luaran (Output)
1	Akuisisi dan Eksplorasi Data	Unduh dataset HCV, meninjau struktur data, dan menganalisis deskripsi statistik awal	Dataset & notebook eksplorasi.
2	Preprocessing & Feature Engineering	Penanganan missing values, encoding, normalisasi, dan pemisahan train-test.	Dataset bersih dan siap dilatih.
3	Implementasi Model	Pelatihan model Logistic Regression dan Random Forest menggunakan dataset bersih.	Dua model ML terlatih.
4	Evaluasi dan Analisis	Menghitung Accuracy, Precision, Recall, F1-Score, serta membuat Confusion Matrix dan analisis fitur.	Grafik hasil evaluasi & laporan analisis.

5	Dokumentasi Akhir	Penyusunan laporan akhir dan pembuatan repositori GitHub.	Laporan akhir & presentasi proyek.

BAB 4: LUARAN PROYEK

Luaran yang akan diserahkan untuk tugas proyek ini meliputi:

1. **Proposal:** Dokumen ini sebagai rencana penelitian dan implementasi proyek.
2. **GitHub Repository:** Berisi kode Python/Jupyter Notebook lengkap mulai dari preprocessing hingga evaluasi model, beserta README.md yang informatif.
3. **Laporan Akhir:** Dokumentasi hasil eksperimen dan analisis perbandingan model.
4. **Link Dataset & Referensi Paper:** Dicantumkan pada bagian referensi.

BAB 5: KESIMPULAN DAN HARAPAN

Dataset HCV Data memberikan peluang besar untuk memanfaatkan teknik machine learning dalam analisis data medis. Dengan membandingkan model linear (Logistic Regression) dan model ensemble (Random Forest), diharapkan proyek ini dapat:

- Menunjukkan algoritma yang paling efektif dalam klasifikasi tahap penyakit hati.
- Memberikan pemahaman tentang fitur laboratorium yang paling signifikan.
- Menjadi dasar bagi pengembangan sistem pendukung keputusan klinis berbasis data.

Langkah selanjutnya meliputi implementasi model, penyempurnaan parameter, dan pembuatan laporan analisis komprehensif terhadap hasil evaluasi.

Referensi

- Smith J., & Doe A. 2023. Machine learning dalam diagnosis hepatitis C: tinjauan sistematis. *Jurnal Informatika Medis*, 15(2): 123-135. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10394001/.
- Hoffmann, G., Bietenbeck, A., Lichtinghagen, R., & Klawonn, F. (2018). *Using Machine Learning Techniques to Generate Laboratory Diagnostic Pathways — A Case Study.* Journal of Laboratory and Precision Medicine.