Proposal

Heart Failure Mortality Prediction Using a Logistic Regression Model

Dosen Pengampu:

Hendri Kharisma S.Kom, M.T

Disusun oleh:

Muhamad Naufal Fauzan	241572010008
Siti Tahtia Ainun Zahra	241572010014

SISTEM INFORMASI

SEKOLAH TINGGI MANAJEMEN INFORMATIKA KOMPUTER (STMIK) TAZKIA

2025

BAB 1: PENDAHULUAN

1.1 Latar Belakang Masalah

Penyakit kardiovaskular (CVDs) merupakan penyebab kematian utama secara global. Salah satu manifestasi utamanya adalah Gagal Jantung (Heart Failure/HF), sebuah kondisi kronis di mana jantung tidak dapat memompa darah secara efisien untuk memenuhi kebutuhan tubuh.

Prediksi prognosis, khususnya mortalitas (kematian), pada pasien HF sangat penting untuk menentukan strategi perawatan dan prioritas intervensi klinis. Rekam medis elektronik (EHR) modern menyediakan data klinis yang kaya untuk dianalisis.

Dengan memanfaatkan teknik *supervised machine learning*, kita dapat membangun model prediktif untuk mengidentifikasi pasien dengan risiko mortalitas tertinggi. Dataset "Heart Failure Clinical Records" telah berhasil digunakan dalam penelitian sebelumnya, seperti dalam sebuah skripsi (Universitas Sriwijaya, 2021) yang menerapkan algoritma **Naive Bayes** untuk kasus yang sama. Dataset ini menyediakan 12 fitur klinis dan 1 target mortalitas, menjadikannya dataset yang ideal untuk riset mendalam menggunakan model klasifikasi **Logistic Regression**.

1.2 Rumusan Masalah

Bagaimana membangun dan menganalisis model *supervised learning* yang akurat menggunakan algoritma **Logistic Regression** untuk memprediksi mortalitas (DEATH_EVENT) pada pasien gagal jantung?

1.3 Tujuan

1. Melaksanakan salah satu Tugas Besar pada mata kuliah Machine Learning.
2. Membangun dan memvalidasi model **Logistic Regression** untuk prediksi mortalitas pasien gagal jantung.
3. Menganalisis koefisien dan *feature importance* dari model Logistic Regression untuk mengidentifikasi faktor-faktor risiko klinis yang paling berpengaruh terhadap mortalitas.
4. Mengevaluasi performa model secara komprehensif menggunakan metrik evaluasi klasifikasi standar.

1.4 Manfaat

1. Memberikan solusi berbasis data untuk prediksi risiko mortalitas pasien HF yang dapat mendukung keputusan klinis.
2. Meningkatkan pemahaman tentang faktor-faktor klinis yang menjadi prediktor kuat kematian akibat gagal jantung melalui interpretasi model linear.
3. Mengembangkan kompetensi dalam penerapan dan analisis mendalam *supervised learning* (khususnya model linear) pada masalah klasifikasi di bidang kesehatan.

BAB 2: DESKRIPSI DATASET

Dataset bersumber dari:

- **Repository:** UCI Machine Learning Repository
- **Judul:** Heart Failure Clinical Records Dataset (Donasi Tahun 2020)
- **URL:** https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records
- **Lisensi:** Creative Commons Attribution 4.0 International (CC BY 4.0)

2.1 Fitur-Fitur Data

Dataset ini terdiri dari 12 fitur prediktor dan 1 variabel target. Variabel target, DEATH_EVENT, adalah biner (0 = tidak meninggal, 1 = meninggal selama periode *follow-up*), sehingga ini adalah **masalah klasifikasi**.

Fitur	**Deskripsi**	**Tipe**
age	Usia pasien	Numerik
anaemia	Apakah ada pengurangan hemoglobin (0=Tidak, 1=Ya)	Biner
creatinine_phosphokinase	Level enzim CPK di darah	Numerik (mcg/L)
diabetes	Apakah pasien memiliki riwayat diabetes (0=Tidak, 1=Ya)	Biner
ejection_fraction	Persentase darah yang dipompa jantung per detak	Numerik (%)
high_blood_pressure	Apakah pasien memiliki hipertensi (0=Tidak, 1=Ya)	Biner
platelets	Jumlah trombosit di darah	Numerik (kiloplatelets/mL)
serum_creatinine	Level kreatinin di darah (indikator fungsi ginjal)	Numerik (mg/dL)
serum_sodium	Level sodium di darah	Numerik (mEq/L)
sex	Jenis kelamin (0=Wanita, 1=Pria)	Biner
smoking	Apakah pasien perokok (0=Tidak, 1=Ya)	Biner
time	Periode follow-up pasien	Numerik (Hari)

DEATH_EVENT (Target)	**Apakah pasien meninggal (0=Tidak, 1=Ya)**	**Biner**

2.2 Ukuran Dataset

- **Total sampel:** 299 pasien
- **Total fitur:** 12 variabel prediktor
- **Target variable:** Biner (Klasifikasi)
- **Missing values:** Tidak ada (dataset sudah *clean*).
- **Keseimbangan Kelas:** Dataset ini *imbalanced*. Terdapat 203 pasien yang bertahan (DEATH_EVENT=0) dan 96 pasien yang meninggal (DEATH_EVENT=1).

2.3 Karakteristik Dataset

Dataset ini ideal untuk pemodelan klasifikasi biner. Karena semua fitur sudah dalam bentuk numerik (termasuk biner 0/1), tidak diperlukan *encoding* kategorikal yang kompleks. Adanya ketidakseimbangan kelas (*imbalance*) menjadi tantangan yang harus ditangani saat evaluasi model.

BAB 3: METODOLOGI

3.1 Preprocessing

1. **Data Loading & Exploration:**
   - *Load* dataset menggunakan pandas.
   - Melakukan *Exploratory Data Analysis* (EDA) untuk memahami distribusi setiap fitur (misal: histogram) dan korelasi antar fitur (heatmap).
   - Menganalisis distribusi kelas target (DEATH_EVENT) untuk mengkonfirmasi *imbalance*.
2. **Feature Scaling:**
   - Algoritma Logistic Regression sangat sensitif terhadap skala data. Oleh karena itu, fitur-fitur numerik (seperti age, creatinine_phosphokinase, dll.) akan distandarisasi menggunakan StandardScaler dari scikit-learn.
3. **Split Dataset:**
   - Membagi dataset menjadi data latih (*training set*) dan data uji (*testing set*) dengan proporsi 80% *training* dan 20% *testing*.
   - Akan digunakan *stratified splitting* untuk memastikan proporsi kelas DEATH_EVENT (0 dan 1) tetap sama di *training* dan *testing set*.

3.2 Modeling

Algoritma yang akan diimplementasikan adalah:

1. **Logistic Regression:**
   - Ini adalah model linear yang sangat umum digunakan untuk klasifikasi biner, terutama di bidang medis karena interpretasinya yang baik. Riset ini akan berfokus pada

penerapan, evaluasi, dan interpretasi mendalam dari model ini.

Tools & Libraries:

- **Bahasa:** Python 3.x
- **Manipulasi Data:** pandas
- **Operasi Numerik:** numpy
- **Machine Learning:** scikit-learn (untuk model, *splitting*, *scaling*, dan metrik evaluasi)
- **Visualisasi:** matplotlib & seaborn

3.3 Evaluasi Model

Karena dataset ini *imbalanced*, metrik **Accuracy** saja tidak cukup. Kami akan menggunakan metrik evaluasi klasifikasi yang komprehensif untuk mengevaluasi model:

1. **Confusion Matrix:** Matriks 2x2 yang menunjukkan True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN).
2. **Accuracy:** (TP + TN) / Total - Persentase prediksi yang benar secara keseluruhan.
3. **Precision:** TP / (TP + FP) - Dari semua yang diprediksi "Meninggal", berapa persen yang benar?
4. **Recall (Sensitivity):** TP / (TP + FN) - Dari semua yang *sebenarnya* "Meninggal", berapa persen yang berhasil terdeteksi? (Sangat penting dalam kasus medis).
5. **F1-Score:** Rata-rata harmonik dari Precision dan Recall. Metrik yang baik untuk data *imbalanced*.
6. **ROC-AUC Score:** Area di bawah kurva ROC. Mengukur kemampuan model secara keseluruhan untuk membedakan antara kelas 0 dan 1.

3.4 Visualisasi

1. **Heatmap Korelasi:** Untuk melihat hubungan antar fitur.
2. **Visualisasi Koefisien (Coefficients):** Dari model Logistic Regression untuk menginterpretasi fitur mana yang paling berpengaruh (positif atau negatif) terhadap mortalitas.
3. **Confusion Matrix Plot:** Visualisasi *heatmap* dari *confusion matrix* model.
4. **ROC Curve Plot:** Menampilkan kurva ROC dari model Logistic Regression.

BAB 4: RENCANA KERJA

4.1 Timeline

Minggu	Aktivitas
1	Studi literatur (paper referensi), *data loading*, EDA, dan *preprocessing* data.
2	Implementasi dan *training* awal model Logistic Regression.

3	*Hyperparameter tuning* (jika diperlukan), finalisasi model, dan analisis koefisien/fitur.
4	Evaluasi akhir model, pembuatan visualisasi, dan penyusunan laporan akhir.

4.2 Deliverables

1. **GitHub Repository:** Berisi kode lengkap dalam format *notebook* (.ipynb) atau *script* (.py), serta file README.md.
2. **Proposal:** Dokumen ini.
3. **Laporan Akhir:** Laporan lengkap yang mendokumentasikan seluruh proses, hasil eksperimen, analisis mendalam, dan kesimpulan.
4. **Presentasi:** Slide presentasi yang merangkum temuan proyek.

BAB 5: KESIMPULAN DAN HARAPAN

Dataset "Heart Failure Clinical Records" menyediakan data klinis yang relevan dan modern (tahun 2020) untuk membangun model prediksi mortalitas pasien. Dengan menerapkan dan menganalisis secara mendalam model Logistic Regression, penelitian ini diharapkan dapat:

1. Menganalisis efektivitas dan performa model Logistic Regression untuk memprediksi mortalitas pada dataset ini.
2. Memberikan *insight* mengenai faktor-faktor klinis apa saja yang paling berpengaruh dalam menentukan prognosis pasien gagal jantung (melalui analisis koefisien Logistic Regression).
3. Mengembangkan model dasar yang dapat dijadikan sebagai alat bantu pendukung keputusan (*decision support tool*) bagi tenaga medis.

Selanjutnya, akan dilakukan implementasi, *hyperparameter tuning*, dan analisis komprehensif terhadap hasil evaluasi model.

Referensi

1. (2021). *Klasifikasi pasien gagal jantung menggunakan metode naive bayes dengan penerapan diskritisasi* [Skripsi, Universitas Sriwijaya]. Repository Universitas Sriwijaya. https://repository.unsri.ac.id/64869/
2. Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M., & Raza, M. A. (2017). *Heart Failure Clinical Records Dataset* (v1). UCI Machine Learning Repository. https://doi.org/10.24432/C5Z89R