Proposal Tugas Besar Mata Kuliah Dasar Kecerdasan Artificial: Regresi pada Dataset *Abalone* menggunakan k-Nearest Neighbors

Disusun oleh:

Muhammad Irham Zidny - 1301223461

PROGRAM STUDI S1 INFORMATIKA

FAKULTAS INFORMATIKA

TELKOM UNIVERSITY

2025

Daftar Isi

1. Pendahuluan

Abalone adalah jenis moluska laut yang memiliki nilai ekonomi tinggi, terutama di sektor perikanan dan kuliner. Usia abalone, yang ditentukan dari jumlah cincin pada cangkangnya, menjadi informasi penting untuk pengelolaan populasi dan konservasi. Proyek ini bertujuan membangun model machine learning berbasis regresi menggunakan algoritma k-Nearest Neighbors (k-NN) untuk memprediksi jumlah cincin abalone berdasarkan ukuran fisik dan beratnya. Dataset yang digunakan berasal dari UCI Machine Learning Repository, yang dikenal cukup bersih dan siap pakai.

2. Latar Belakang

Penentuan usia abalone secara manual melalui penghitungan cincin cangkang memakan waktu dan tenaga. Dengan machine learning, prediksi usia dapat dilakukan lebih cepat menggunakan fitur-fitur fisik seperti panjang, diameter, dan berat. Pendekatan ini mendukung nelayan dan ilmuwan dalam mengelola sumber daya abalone secara berkelanjutan. Algoritma k-NN dipilih karena sifatnya yang sederhana namun efektif untuk menangani data numerik dengan hubungan non-linier, cocok untuk kasus regresi pada dataset ini.

3. Dataset

Dataset Abalone dari UCI Machine Learning Repository berisi 4.177 sampel dengan 9 fitur, termasuk satu kolom target (jumlah cincin). Data ini sudah bersih, tanpa nilai hilang, sehingga tidak memerlukan imputasi atau pembersihan tambahan. Namun, preprocessing tetap dilakukan, yaitu:

- **Encoding**: Kolom kelamin (kategorikal) diubah menjadi numerik menggunakan one-hot encoding.
- **Normalisasi**: Fitur-fitur numerik diskalakan ke rentang yang sama untuk menghindari bias akibat perbedaan skala.
- **Pembagian data**: Data dibagi menjadi 80% data latih dan 20% data uji untuk evaluasi model.

Link dataset: https://archive.ics.uci.edu/dataset/1/abalone

4. Deskripsi Fitur dan Permasalahannya

Fitur dalam dataset meliputi:

1. **Kelamin** (M, F, I): Kategori jenis kelamin abalone (jantan, betina, atau infantil).
2. **Panjang** (mm): Panjang cangkang abalone.
3. **Diameter** (mm): Diameter cangkang abalone.
4. **Tinggi** (mm): Tinggi cangkang abalone.
5. **Berat keseluruhan** (gram): Berat total abalone.
6. **Berat daging** (gram): Berat daging abalone.
7. **Berat usus** (gram): Berat organ dalam setelah pendarahan.
8. **Berat cangkang** (gram): Berat cangkang kering.
9. **Jumlah cincin**: Target, menunjukkan usia abalone (cincin + 1,5 ≈ usia dalam tahun).

5. Permasalahan:

- Kolom kelamin bersifat kategorikal, sehingga perlu diencode agar dapat digunakan dalam model k-NN.
- Perbedaan skala antar fitur (misalnya, berat dalam gram vs panjang dalam mm) dapat memengaruhi performa k-NN, yang sensitif terhadap jarak. Normalisasi diperlukan untuk menyeragamkan skala.
- Jumlah cincin sebagai target bersifat kontinu (meskipun diskrit dalam praktik), sehingga k-NN harus diatur untuk regresi, bukan klasifikasi.
- Dataset relatif kecil, sehingga pemilihan parameter k pada k-NN dan proporsi data latih-uji perlu dioptimalkan untuk mencegah overfitting atau underfitting.