

Kinerja Diskritisasi Metode Binary Encoding Equal Width dan Equal Frequency Terhadap Fitur Data pada Klasifikasi

Pramaishella Ardiani¹, Sri Suryani Prasetyowati², Yuliant Sibaroni³

^{1,2,3}Fakultas Informatika, Universitas Telkom, Bandung

⁴Divisi Digital Service PT Telekomunikasi Indonesia

¹shellarp@students.telkomuniversity.ac.id, ²srisuryani@telkomuniversity.ac.id,

³yuliant@telkomuniversity.ac.id,

Abstrak

Proses klasifikasi sering kali memerlukan berbagai proses untuk meningkatkan tingkat akurasi, yang dapat disebabkan oleh berbagai penyebab, seperti rentang nilai atribut yang terlalu lebar pada dataset. Pendekatan diskritisasi memberikan solusi yang efektif untuk mengatasi masalah ini. Penelitian ini bertujuan untuk mengevaluasi dan membandingkan efektivitas dua pendekatan diskritisasi, yaitu Equal-Width dan Equal-Frequency, dalam meningkatkan akurasi algoritma klasifikasi. Analisis akan dilakukan pada dataset dengan ukuran yang berbeda. Penelitian ini menggunakan model klasifikasi Naïve Bayes, Decision Tree, dan Support Vector Machine, dengan tiga dataset: data Kemacetan Lalu Lintas Kota Bandung (3804 data), data kasus COVID-19 Kota Bandung (2718 data), dan data Penyakit Demam Berdarah Kota Bandung (150 data). Tiga skenario percobaan dilakukan untuk mengevaluasi pengaruh kedua metode diskritisasi terhadap akurasi. Skenario awal tidak mengimplementasikan diskritisasi, sementara skenario kedua menggunakan metode diskritisasi Equal-Width, dan skenario ketiga menggunakan diskritisasi Equal-Frequency. Hasil akhir penelitian menunjukkan peningkatan yang signifikan dalam akurasi setelah proses diskritisasi. Ketika diterapkan pada dataset Lalu Lintas, model Naïve Bayes menunjukkan tingkat akurasi 94%. Sementara itu, model Decision Tree menghasilkan tingkat akurasi 71% untuk dataset COVID-19 dan tingkat akurasi luar biasa sebesar 98% untuk dataset Penyakit Demam Berdarah. Hasil di atas menunjukkan bahwa penggunaan teknik diskritisasi Equal-Width dan Equal-Frequency secara efektif mengatasi masalah rentang nilai atribut yang terlalu lebar dalam prosedur klasifikasi.

Kata kunci : *Equal-Width, Equal-Frequency, Diskritisasi, Klasifikasi, Akurasi*

Abstract

The classification process frequently requires assistance in improving the accuracy levels, which can be ascribed to many causes, such as the dataset's wide range of attribute values. Discretization approaches provide a viable approach to mitigate these concerns. This study aims to evaluate and compare the efficacy of two discretization approaches, namely Equal-Width and Equal-Frequency, in improving the accuracy of classification algorithms. The analysis will be conducted on datasets of different sizes. The study utilizes Naïve Bayes, Decision Tree, and Support Vector Machine as classification models, employing three datasets: Bandung City Traffic data (3804 recordings), Bandung City COVID-19 cases data (2718 records), and Bandung City Dengue Fever Disease Index data (150 records). Three experimental scenarios were conducted to evaluate the influence of the two discretization approaches on accuracy. The initial scenario does not involve any form of discretization. In contrast, the second situation utilizes the Equal-Width method, while the third scenario employs the Equal-Frequency discretization approach. The experimental findings demonstrate a notable increase in accuracy following the discretization process. When applied to the Traffic dataset, the Naïve Bayes model demonstrated a 94% accuracy rate. In contrast, the Decision Tree model yielded a 71% accuracy rate for the COVID-19 dataset and an outstanding 98% accuracy rate for the Dengue Fever Disease dataset. The results above indicate that using EqualWidth and Equal-Frequency discretization techniques effectively tackles the issue of wide attribute value ranges during the classification procedure.

Keywords: *Equal-Width, Equal-Frequency, Discretization, Classification, Accuracy*

1. Pendahuluan [10 pts/Bold]

Latar Belakang

Persiapan data merupakan salah satu tahapan yang diperlukan dalam data mining (DM) sebelum menerapkan algoritma klasifikasi data mining (DMCA) pada data. Pra-pemrosesan data adalah fase penting dalam penambangan data yang mencakup teknik seperti transformasi data, pembersihan, reduksi, dan diskritisasi [1]. Variasi nilai data yang diolah dalam data mining sering ditemukan pada satu atribut antara nilai yang satu dengan nilai yang lain mempunyai rentang atau gap yang terlalu jauh. Dalam beberapa dekade terakhir, para