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

Pramaishella Ardiani[1], Sri Suryani Prasetiyowati[2], Yuliant Sibaroni[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[4]Divisi Digital Service PT Telekomunikasi Indonesia
[1]shellarp@students.telkomuniversity.ac.id, [2]srisuryani@telkomuniversity.ac.id,
[3]yuliant@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

peneliti telah mempelajari salah satu algoritma diskritisasi yang kini menjadi salah satu teknik preprocessing yang paling sering digunakan dalam data mining [2].

Prinsip kerja diskritisasi adalah mengubah atau memisahkan sifat-sifat kontinu ke dalam kategori-kategori atau nominal. Variabel dengan nilai kontinu diubah menjadi variabel diskrit, dibangun dengan beberapa interval yang tidak tumpang tindih dalam proses ini. Manfaat utama diskritisasi adalah: (1) data berkurang dan menggunakan lebih sedikit ruang penyimpanan. (2) Data diskrit lebih mudah dipahami, dimanfaatkan, dan dijelaskan karena lebih mendekati tingkat pengetahuan dibandingkan data kontinu. (3) Dengan data diskrit, DMCA dapat bekerja lebih cepat dan mencapai akurasi klasifikasi yang lebih tinggi. Algoritma klasifikasi data mining seperti Naïve Bayes, Decision Tree, dan Support Vector Machine adalah sepuluh model klasifikasi yang sering digunakan [3]. Namun algoritma tersebut hanya dapat memproses atau mengolah data numerik. Hal ini bertujuan untuk memastikan bahwa dengan menggunakan diskritisasi, kinerja akan dimaksimalkan dengan menciptakan model yang lebih efektif dan efisien [4].

Penelitian yang telah dipelajari oleh [2] membandingkan algoritma diskritisasi dalam klasifikasi pohon keputusan. Selain itu, penelitian [5] menerapkan diskritisasi tanpa pengawasan dalam klasifikasi naif Bayes. Pada penelitian yang dilakukan [6], penggunaan diskritisasi equal-width dengan model klasifikasi Naive Bayes meningkatkan nilai akurasi pada pasien TB hingga 81%. Ketiga penelitian tersebut membuktikan bahwa diskritisasi pada preprocessing data dapat meningkatkan nilai akurasi suatu model klasifikasi.

Berdasarkan uraian mengenai diskritisasi metode unsupervised equal-width dan equal-frekuensi pada penelitian-penelitian sebelumnya. Penelitian ini kemudian akan dibedakan menjadi tiga skenario: tanpa preprocessing atau diskritisasi data, penerapan diskritisasi equal-width, dan penerapan diskritisasi frekuensi setara. Penelitian ini bertujuan untuk menerapkan metode diskritisasi unsupervised equal-width dan equal-frekuensi pada fitur data dengan rentang yang terlalu jauh dapat meningkatkan model klasifikasi. Kontribusi penelitian ini adalah untuk mengidentifikasi pengaruh jumlah data terhadap kinerja diskritisasi equal-width dan equal-frekuensi tanpa pengawasan.

Topik dan Batasannya

Berdasarkan latar belakang dan identifikasi masalah yang telah dijelaskan, dapat dirumuskan masalah dalam penelitin ini sebagai berikut:

1. Bagaimana proses dikristisasi equal-width dan equal-frequency pada dataset penyakit covid-19, dataset penyakit demam berdarah dan dataset arus lalu lintas di kota Bandung?
2. Bagaimana implementasi naïve bayes, decision tree dan Support Vector Machine dalam implementasi metode diskritisasi?
3. Bagaimana kinerja hasil diskritasasi antara equal-width dan equal-frequency saat menggunakan algoritma Naïve Bayes, Decision Tree dan Support Vector Machine?

Tujuan

Tujuan yang ingin dicapai pada penelitian ini adalah:

1. Untuk mengetahui proses diskritisasi menggunakan metode equal-width dan equal-frequency pada dataset penyakit covid, dataset penyakit demam berdarah dan dataset arus lalu lintas di kota Bandung.
2. Untuk mengetahui implementasi algoritma Naïve Bayes, Decision Tree dan Support Vector Machine dalam proses klasifikasi menggunakan metode diskritisasi.
3. Untuk mengetahui performa masing-masing metode equal-width dan equal-frequency pada algoritma Naïve Bayes, Decision Tree dan Support Vector Machine.

2. Studi Terkait

Diskritisasi adalah teknik pra-pemrosesan data yang mengubah data kontinu menjadi nilai diskrit. Metode ini dapat berguna untuk berbagai tugas, seperti klasifikasi, pengelompokan, dan regresi. Terdapat banyak metode diskritisasi yang berbeda, masing-masing dengan kelebihan dan kekurangannya. Dua metode diskritisasi yang paling umum adalah dengan lebar yang sama dan frekuensi yang sama. Diskritisasi dengan lebar yang sama membagi rentang data kontinu menjadi sejumlah interval tetap dengan lebar yang sama. Diskritisasi frekuensi yang sama membagi rentang data kontinu menjadi sejumlah interval tetap dengan frekuensi yang sama.

Banyak peneliti telah meneliti bagaimana kinerja algoritma diskritisasi pada masalah klasifikasi [2]. Melakukan penelitian yang membandingkan algoritma diskritisasi dalam klasifikasi pohon keputusan. Selanjutnya peneliti [5] menggunakan diskritisasi tanpa pengawasan dalam klasifikasi Naïve Bayes. Pada penelitian yang dilakukan oleh [6] menggunakan diskritisasi equal-width dengan model klasifikasi Naïve Bayes, nilai akurasi pada pasien TB meningkat menjadi 81%. Ketiga penelitian tersebut membuktikan