BAB I

PENDAHULUAN

I.1 Latar Belakang

Data mining merupakan suatu bidang yang sangat berkembang pesat karena adanya kebutuhan untuk memanfaatkan *data warehouse* yang dimiliki, yaitu dengan melakukan serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual. Ada beberapa tantangan dalam data mining salah satunya fenomena *curse of dimensionality*. Fenomena *curse of dimensionality* merupakan fenomena yang terjadi pada data mining dengan adanya peningkatan jumlah dimensi data sehingga dapat menyebabkan turunnya performansi. Oleh karena itu harus dilakukan *preprocessing data* yaitu dengan *data reduction* (Kantardzic, 2011).

Data reduction sangat membantu mereduksi data yang kompleks tanpa mengurangi integritas dari data yang asli dan tidak mengurangi kualitas informasi yang dihasilkan. Teknik reduksi data dapat diterapkan untuk memperoleh representasi berkurangnya jumlah kumpulan data yang volumenya jauh lebih kecil, namun dengan tetap mempertahankan integritas data asli. Sehingga data set yang dikurangi harus lebih efisien dan menghasilkan analisis yang sama atau hampir sama (Han et al., 2012).

Strategi untuk *data reduction* meliputi:

1. *Data cube aggregation*, di mana operasi agregasi diterapkan pada data dalam konstruksi kubus data.

2. *Attribute subset selection*, di mana atribut atau dimensi yang tidak relevan, lemah relevan, atau berlebihan dapat dideteksi dan dihapus.
3. *Dimensionality reduction*, di mana mekanisme pengkodean digunakan untuk mengurangi ukuran kumpulan data.
4. *Numerosity reduction*, di mana data diganti atau diperkirakan dengan alternatif, representasi data yang lebih kecil seperti model parametrik (yang hanya perlu menyimpan parameter model daripada data aktual) atau metode nonparametric seperti pengelompokan, pengambilan sampel, dan penggunaan histogram.
5. *Discretization and concept hierarchy generation*, di mana nilai data mentah untuk atribut digantikan oleh rentang atau tingkat konseptual yang lebih tinggi. Diskritisasi data adalah bentuk pengurangan angka yang sangat berguna untuk pembuatan hierarki konsep secara otomatis. Diskretisasi dan pembuatan hierarki konsep adalah alat yang ampuh untuk penambangan data, karena memungkinkan penambangan data pada berbagai tingkat abstraksi.

Penelitian ini akan mempergunakan *dimensionality reduction* yang nantinya akan digunakan untuk mereduksi dimensi pada *dataset* yang berdimensi banyak, *dimensionality reduction* dibagi menjadi dua bagian, yaitu:

1. *Feature Extraction* atau ekstraksi fitur adalah istilah umum untuk metode membangun kombinasi variabel untuk mengatasi masalah ini sambil tetap menggambarkan data dengan akurasi yang cukup. Banyak praktisi

pembelajaran mesin percaya bahwa ekstraksi fitur yang dioptimalkan dengan benar adalah kunci untuk konstruksi model yang efektif.

2. *Feature Selection* atau seleksi fitur merupakan tahap pertama dalam proses klasifikasi. Data nyata mungkin berisi fitur dari berbagai relevansi untuk memprediksi label kelas. Fitur yang tidak relevan biasanya akan merusak keakuratan model klasifikasi selain menjadi sumber inefisiensi komputasi. Oleh karena itu, tujuan dari algoritma pemilihan fitur adalah untuk memilih fitur yang paling informatif sehubungan dengan label kelas (Aggarwal, 2015). Tiga jenis utama metode yang digunakan untuk seleksi fitur dalam klasifikasi:

    a. *Filter Approach*.

        Tersedia kriteria matematis yang tajam untuk mengevaluasi kualitas fitur atau subset fitur. Kriteria ini kemudian digunakan untuk menyaring fitur yang tidak relevan.

    b. *Wrapper Approach.*

        Diasumsikan bahwa algoritma klasifikasi tersedia untuk mengevaluasi seberapa baik kinerja algoritma dengan subset fitur tertentu. Sebuah algoritma pencarian fitur kemudian membungkus algoritma ini untuk menentukan himpunan yang relevan dari fitur.

    c. *Embedded Approach*.

        Solusi untuk model klasifikasi sering kali berisi petunjuk yang berguna tentang fitur yang paling relevan. Fitur seperti itu diisolasi,

dan pengklasifikasinya adalah dilatih ulang pada fitur yang dipangkas.

Pada penelitian ini metode klasifikasi *decision tree* digunakan untuk melakukan penilaian performa *data reduction* pada *dataset* yang berdimensi banyak. Metode klasifikasi *decision tree* digunakan karena metode ini efisien jika digunakan pada data yang dimensi sedikit (mungkin kurang dari 10) akan tetapi gagal memberikan hasil yang berarti ketika digunakan pada data yang jumlah dimensinya banyak (Rokach & Maimon, 2015).

Beberapa penelitian sebelumnya telah melakukan perbandingan beberapa metode *data reduction* pada *dataset* yang berdimensi banyak. Seperti yang dilakukan oleh (Kavitha & Kannan, 2016) mengusulkan *feature extraction* dengan teknik *principal component analysis* (PCA) dan *feature selection* menggunakan *wrapper approach* serta kriteria *split* dari algoritma C4.5 untuk digunakan sebagai metode untuk menghasilkan fitur-fitur yang relevan dari *dataset* UCI *heart dataset*. Metode yang diusulkan dapat menghasilkan fitur-fitur yang relevan sehingga meningkatkan efisiensi dan akurasi.

(Yadav & Sehra, 2018) mengusulkan skema *watermarking* baru menggunakan *feature extraction* dengan teknik *principal component analysis* (PCA) dan *singular value decomposition* (SVD) pada *dual tree complex wavelet transforms* (DTCWT) di *low dimensional subspace* untuk gambar *grayscale*. Menghasilkan nilai *peak signal to noise ratio* (PSNR) dan korelasi yang tinggi dibandingkan dengan *state of art techniques* menunjukkan ketidakterlihatan dan ketahanan skema pada metode yang diusulkan.

(Bahassine et al., 2020) mengusulkan *feature selection* dengan teknik *chi-squared* pada klasifikasi *support vector machine* (SVM) untuk *text mining*. Metode yang diusulkan terbukti secara signifikan meningkatkan kinerja model klasifikasi teks arab.

(Sallehuddin et al., 2021) mengusulkan metode pemilihan fitur *ensemble multi filter* yaitu *information gain, gain ratio, chi-square* dan *relief-f*, kemudian *support vector machine* digunakan untuk mengevaluasi kinerja klasifikasi fitur yang dipilih. Setelah fitur *ensemble* didapatkan, selanjutnya dilakukan optimasi dengan parameter klasifikasi secara sinkron dengan *particle swarm optimization* dan *support vector machine*. Pengujian menggunakan *dataset breast cancer* dan l*ymphography* dari UCI *machine learning repository*. Hasil dari eksperimen menunjukkan bahwa metode yang diusulkan berhasil menunjukkan kinerja akurasi *classifier* dengan optimal fitur yang signifikan dibandingkan dengan metode lain yang ada seperti PSO-SVM dan *classical* SVM. Oleh karena itu, metode yang diusulkan dapat digunakan sebagai metode alternatif untuk menentukan solusi optimal dalam menangani data yang berdimensi tinggi.

(Zhang et al., 2019) mengusulkan *feature selection* dengan teknik l*arge margin hybrid algorithm for feature selection* (LMFS) yang merupakan *hybrid filter/wrapper algorithm*, pada klasifikasi menggunakan k-NN, PLS-DA dan LS-SVM untuk *dataset coffees, meats, two tablets*, *olive oils* dan *tobaccos* dari *quadram institute dataset*. Fitur yang dipilih oleh metode LMFS memiliki kinerja klasifikasi dan interpretasi model yang lebih baik. Selain itu, LMFS dapat secara efektif mengatasi dampak kompleksitas pengklasifikasi pada waktu komputasi dan

distance-based classifiers ditemukan lebih cocok untuk memilih subset terakhir dalam LMFS.

(Silva et al., 2021) mengusulkan *feature selection* dengan teknik *forward selection* dan *backward selection* pada implementasi klasifikasi *pearson's linear correlation*, *relieff*, *welch's t-test* dan algoritma berbasis *multilinear regression* untuk pemilihan *feature acoustic* identifikasi patologi suara. Dari metode-metode yang diusulkan didapat bahwa metode *relieff algorithm* dengan *feature selection* dapat memberikan performa terbaik dibanding metode lainnya.

(Bhat & Dutta, 2021) mengusulkan *feature discrimination* dan *information gain* pada pemilihan fitur *multi-tier model*, yang dapat menemukan fitur yang relevan dan signifikan untuk meningkatkan akurasi pendeteksian *malware*. Kemudian kumpulan fitur yang dipilih diaplikasikan pada lima teknik klasifikasi *machine learning*. Pengujian dilakukan pada *dataset Virustotal, VirusShare* dan *Drebin* yang diperoleh dari pihak ketiga. Pengujian dan analisis yang ketat menunjukkan bahwa klasifikasi *Random Forest* mencapai tingkat akurasi tertinggi sebesar 96,28%.

(Pasha & Mohamed, 2022) mengusulkan metode *advanced hybrid ensemble gain ratio feature selection* (AHEG-FS) pada *machine learning* untuk meningkatkan prediksi risiko penyakit jantung, teknik pemilihan fitur yang digunakan yaitu: *ensemble feature selection*, *gain ratio feature selection* dan *backward feature elimination*, kemudian menggunakan *area under the curve* (AUC) sebagai evaluasi untuk prediksi risiko penyakit yang efektif. Pengujian dilakukan pada 4 *dataset Irvine M*L *repository* dari *University of California* yaitu:

heart disease datasets Cleveland, *Hungarian*, *Statlog* dan *Switzerland*. Dari penelitian didapati bahwa *backward feature elimination* menghasilkan perolehan subset terbaik dari fitur yang sangat berkontribusi dan menghasilkan presisi tertinggi.

(Kshirsagar & Kumar, 2021) mengusulkan metode reduksi fitur berdasarkan kombinasi algoritma reduksi fitur berbasis filter, yaitu *information gain ratio* (IGR), *correlation* (CR), dan *relieff* (ReF) pada *dataset* CICIDS 2017 DoS dan KDD *Cup* 99. Tujuan penelitian ini adalah untuk peningkatan kinerja dengan mengurangi jumlah fitur untuk mendeteksi serangan DoS. Dari penelitian didapati bahwa pengurangan fitur yang diusulkan dapat mengurangi jumlah fitur CICIDS 2017 dan KDD *Cup* 99 dari 77 menjadi 24, dan 41 menjadi 12, menghasilkan peningkatan akurasi 99,9593% menggunakan pengklasifikasi *projective adaptive resonance theory* (PART) dengan waktu proses 133,66 s pada *dataset* CICIDS 2017. metode ini juga menghasilkan akurasi yang signifikan pada *dataset* KDD *Cup* 99.

(Gárate-Escamila et al., 2020) mengusulkan metode *dimensionality reduction chi-square*, *principal component analysis* (PCA) dan *chi-square* dan *principal component analysis* (CHI-PCA) dengan tujuan untuk menemukan ciri penyakit jantung dengan menerapkan teknik seleksi ciri pada *dataset heart disease* dari UCI *machine learning repository*, kemudian divalidasi menggunakan enam jenis pengklasifikasian pada *machine learning*. Metode *chi-square* dan *principal component analysis* (CHI-PCA) dengan *random forests* (RF) memiliki akurasi

tertinggi, dengan 98,7% untuk *dataset Cleveland*, 99,0% untuk *dataset Hungaria*, dan 99,4% untuk *dataset Cleveland-Hungaria* (CH).

(Nimbalkar & Kshirsagar, 2021) metode yang diusulkan adalah pemilihan fitur untuk sistem deteksi intrusi (IDS) menggunakan *information gain* (IG) dan *gain ratio* (GR) dengan peringkat 50% fitur teratas untuk mendeteksi serangan DoS dan DDoS, kemudian dievaluasi dan divalidasi dengan *JRipclassifier*. Pengujian menggunakan *dataset* IoT-BoT dan KDD *Cup* 1999. Sistem mencapai akurasi dan tingkat deteksi yang lebih tinggi masing-masing 99,9993%, dan 99,5798%, dengan *JRip* menggunakan 16 fitur pada *dataset* BoT-IoT. Validasi sistem *dataset* KDD *Cup* 1999 juga meningkatkan akurasi dan tingkat deteksi 99,9920% dan 99,9943% untuk mendeteksi serangan DoS menggunakan 19 fitur dengan *JRip*.

Berdasarkan latar belakang diatas, maka penulis membandingkan beberapa metode *data reduction* pada *dataset* yang berdimensi banyak. Kemudian untuk melakukan penilaian performa akan digunakan metode klasifikasi *decision tree* dan kemudian menguji apakah ada perbedaan akurasi yang signifikan antara beberapa metode *data reduction* menggunakan pengujian *t-test*.

I.2 Rumusan Masalah

Berdasarkan latar belakang diatas, maka rumusan masalah dalam penelitian ini adalah bagaimana perbandingan performa beberapa metode *data reduction* pada *dataset* yang berdimensi banyak menggunakan metode klasifikasi *decision tree*. *Performance* yang akan dibandingkan adalah:

1. Tingkat akurasi.

2. Lama waktu proses.
3. Adakah perbedaan akurasi yang signifikan antara metode *data reduction.*

I.3 Tujuan

Penelitian ini dilakukan dengan tujuan untuk mencari metode *data reduction* mana yang dapat memberikan performa yang lebih baik jika digunakan pada *dataset* yang berdimensi banyak menggunakan metode klasifikasi *decision tree.*

I.4 Manfaat

Manfaat yang diharapkan dari hasil penelitian ini adalah sebagai berikut:

1. Memberikan informasi mendalam terhadap perbandingan performa beberapa metode *data reduction* yang digunakan pada *dataset* yang berdimensi banyak menggunakan metode klasifikasi *decision tree*.
2. Menjadi referensi untuk penelitian-penelitian setopik lainnya.

I.5 Batasan Masalah

Berdasarkan perumusan masalah diatas terdapat batasan masalah sebagai berikut:

1. Penelitian ini menggunakan beberapa metode *data reduction* khususnya *dimension reduction*, yaitu *principal component analysis* (PCA), *singular value decomposition* (SVD), *information gain* (IG), *gain ratio* (GR), *chi-square* (CS), *forward selection* (FS) dan *backward elimination* (BE).

2. Penelitian ini menggunakan metode klasifikasi *decision tree* untuk melakukan penilaian performa dan *t-test* untuk melihat apakah ada perbedaan akurasi yang signifikan antara metode *data reduction.*
3. Penelitian ini menggunakan *public dataset* UCI *glass identification* yang berasal dari (https://archive.ics.uci.edu/ml/datasets/glass+identification).
4. Penelitian ini menggunakan aplikasi Rapid Miner® untuk membantu dalam mendapatkan hasil pada proses *data reduction,* validasi, pengujian, evaluasi model klasifikasi *decision tree* dan *t-test*.

I.6 Sistematika Pembahasan

Penyusunan tesis ini menggunakan kerangka pembahasan yang terbentuk dalam susunan bab yang dapat dijelaskan sebagai berikut:

BAB I : PENDAHULUAN

Bab ini berisikan tentang latar belakang masalah, rumusan masalah, batasan masalah, tujuan dan manfaat penelitian serta sistematika penulisan.

BAB II : TINJAUAN PUSTAKA

Bab ini berisikan tentang landasan teori yang mendukung penelitian yang akan dilakukan.

BAB III : METODOLOGI PENELITIAN

Bab ini berisikan tentang data yang digunakan dan metode yang digunakan di dalam pelaksanaan penelitian.

BAB IV : HASIL

Bab ini berisikan hasil pengujian pada metode yang digunakan di dalam pelaksanaan penelitian.

BAB V : PEMBAHASAN

Bab ini berisikan pembahasan dari hasil pengujian yang telah dilakukan.

BAB VI : KESIMPULAN DAN SARAN

Bab ini berisikan kesimpulan dari pembahan dari hasil pengujian yang telah dilakukan dan saran yang dapat dikembangkan dari penelitian ini untuk penelitian selanjutnya.