BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan dalam ilmu data dan penemuan pengetahuan memiliki peran yang sangat penting dalam berbagai aplikasi digital saat ini. Salah satu aspek kunci dalam analisis data adalah proses *feature selection*, yang bertujuan untuk mengurangi dimensi data dan mengatasi biaya komputasi yang tinggi. Proses ini dapat dilakukan melalui dua teknik berbeda, yaitu *feature extraction* dan *feature selection* [1]

Fitur-fitur dalam sebuah dataset seringkali memiliki dimensi yang tinggi, namun hanya sedikit sampel yang memiliki hubungan langsung dengan tugas-tugas penambangan data dan pembelajaran mesin. Hal ini menyebabkan tantangan besar dalam mengekstraksi pola atau informasi yang potensial dari dataset, terutama dalam tugas-tugas penambangan data. Selain itu, kerja dengan dataset yang memiliki dimensi yang tinggi juga meningkatkan kesulitan dalam penemuan pengetahuan dan klasifikasi pola, karena terdapat banyak fitur yang redundan dan tidak relevan [2].

Salah satu metode yang banyak digunakan dalam *feature selection* adalah *information gain* (IG). Metode ini merupakan salah satu metrik yang digunakan untuk menilai seberapa banyak informasi yang diberikan oleh suatu fitur terhadap target atau label yang ingin diprediksi. Dengan menggunakan metode IG, pengguna dapat menentukan fitur yang paling penting untuk dimasukkan ke dalam model. Namun, penggunaan IG sendiri masih terbuka untuk penelitian lebih lanjut, terutama dalam hal pengaturan threshold untuk menentukan signifikansi fitur [3].

Terlepas dari pentingnya *feature selection*, masih ada kesenjangan dalam penelitian sebelumnya terkait dengan penggunaan metode *feature selection* yang efektif, terutama dalam menghadapi dataset yang berdimensi tinggi. Beberapa penelitian telah menunjukkan bahwa penggunaan threshold untuk menilai signifikansi fitur dapat memengaruhi kinerja model. Namun, masih perlu penelitian lebih lanjut untuk mengevaluasi dan membandingkan berbagai metode *feature selection* yang ada untuk memastikan *feature selection* yang optimal [4].

Metode *feature selection* dapat berguna untuk mengurangi jumlah fitur yang redundan dan *noise* dalam sebuah dataset, sebuah proses yang dikenal sebagai

reduksi dimensi (*dimensionality reduction*). Reduksi dimensi bertujuan untuk mengurangi dimensi fitur-fitur dalam sebuah dataset, namun tetap mempertahankan data yang relevan. Dalam dataset yang telah direduksi, fitur-fitur penting tetap dipertahankan meskipun beberapa pola spesifik mungkin menghilang. Pendekatan ini tidak hanya membantu dalam mengurangi ukuran data masukan, namun juga mempertahankan sebagian besar variasi dari fitur-fitur yang penting dibandingkan dengan dataset yang lebih besar. Dengan demikian, data dalam dunia nyata lebih mudah untuk dideteksi dan digunakan dalam aplikasi *data mining*, serta memberikan kinerja akurasi yang tinggi. Selain itu, reduksi dimensi juga bertujuan untuk meningkatkan akurasi dan efisiensi komputasi dalam *data mining*, dan dianggap sebagai tahap pra-pemrosesan yang penting. Selain itu, reduksi dimensi memberikan beberapa keuntungan, seperti menghilangkan pola-pola yang tidak relevan dan redundan dalam dataset, sehingga mengurangi waktu dan jumlah memori yang dibutuhkan untuk memproses data tersebut [5].

Beberapa penelitan telah mengeksplorasi feature selection sebagai fokus penelitiannya. Salah satu contohnya adalah penelitian yang dilakukan oleh Maria I. dkk. menggunakan *Information Gain* dan penggunaan *Threshold* nilai standard deviasi untuk melakukan *feature selection* dan klasifikasi menggunakan algoritma *Random Forest* [6]. Selain itu, penelitian yang dilakukan I Gusti B. dkk. menggunakan metode *Information Gain* untuk melakukan *Feature Selection*, *wrapper techniques* digunakan untuk mengihilangkan atribut yang tidak relevan [1]. Lalu, penelitian yang dilakukan oleh Sri H. dkk. menggunakan *information gain* untuk melakukan *feature selection* dan algoritma *Naive Bayes* untuk melakukan klasifikasi [2].

Kajian penelitian yang dilakukan oleh Maria I. dkk. Pada penelitian ini menggunakan *information gain* untuk melakukan *feature selection*. Lalu menggunakan *threshold* nilai *standard deviation* dari *information gain* yang sudah dihasilkan oleh masing-masing fitur dalam dataset yang digunakan. Terdapat 10 dataset asli yang dilakukan pengujian, lalu ditransformasikan dengan FFT dan IFFT dan selanjutnya diklasifikasi menggunakan algoritma *random forest*. Penelitian ini membandingkan dengan metode *Correlation-Base Feature Selection* (CBFS) dan menggunakan nilai *threshold* sebesar 0.05. Hasil dari penelitian tersebut ketika dataset ditransformasikan dengan FFT dan IFFT didapatkan bahwa nilai akurasi yang diperoleh lebih rendah dan waktu eksekusi yang lebih lama dibandingkan dengan *Correlation-Base Feature selection* (CBFS) dan nilai *threshold* 0.05. Terdapat hasil lain juga yang menunjukkan bahwa dengan menggunakan *original*

dataset dan nilai threshold *standard deviation* memiliki hasil akurasi *feature selection* klasifikasi *random forest* yang lebih baik apabila dibandingkan dengan metode *Correlation-Base Feature selection* (CBFS)[6].

Kajian penelitian yang dilakukan oleh I Gusti B. dkk. *Information gain* digunakan untuk mencari fitur yang tidak relevan dengan label *class* dan menggunakan *wrapper techniques* untuk menghilangkan atribut-atribut yang tidak relevan. Pada penelitian ini menggunakan dataset 1999 NSL-KDD dataset. Dataset tersebut memiliki 41 fitur yang memiliki sifat *continous* dan *discrete* dengan label normal atau anomali. Dataset yang sudah diambil, selanjutnya dilakukan normalisasi dan diskrit. Setelah sudah dilakukan proses tersebut, maka dilakukan proses *feature selection* dengan menggunakan *information gain*. Hasil dari penelitian tersebut dapat disimpulkan bahwa *information gain* dapat digunakan untuk mengetahui pengaruh atribut pada dataset terhadap proses klasifikasi [1].

Kajian penelitian yang dilakukan oleh Sri H. dkk. Algoritma *naive bayes* digunakan untuk melakukan klasifikasi dari data yang sudah diambil. Dataset yang digunakan adalah data kelulusan mahasiswa STMIK YMI Tegal. Terdapat metode *feature selection* yang digunakan yaitu *information gain* yang digunakan secara integrasi agar akurasi dari algoritma yang digunakan dapat mengalami peningkatan. Diambil 8 dari 10 atribut terbaik dari hasil pembobotan setiap *information gain* atribut. Hasil akurasi yang didapatkan dari penggabungan algoritma *naive bayes* dengan *feature selection information gain* yaitu 93,33%. Hasil tersebut lebih besar jika dibandingkan hanya menggunakan algoritma *naive bayes*. Hasil akurasi apabila hanya menggunakan algoritma *naive bayes* yaitu 80,00% [2].

Kajian penelitian yang dilakukan oleh Tomas P. dkk. Algoritma *naive bayes, random forest, decision tree, logistic regression,* dan *logistic regression* digunakan untuk melakukan klasifikasi *text reviews*. Dataset yang digunakan adalah Amazon *customers product-review data for Android Apps*. Terdapat empat tahapan alur kerja dari penelitian ini, yaitu *data extraction, preparation of review texts, bag of words,* dan *classification*. Pada proses klasifikasi digunakan metode *k-fold cross validation* dengan nilai k=10. Hasil akurasi yang didapatkan adalah algoritma *logistic regression* mencapai tingkat akurasi klasifikasi tertinggi (antara 32.43% hingga 58.50%) dalam klasifikasi ulasan produk jika dibandingkan dengan metode klasifikasi *naive bayes*, *random forest, decision tree*, dan *support vector machines*. Sebaliknya, *Decision Tree* memiliki nilai akurasi rata-rata terendah (dari 24.10% hingga 34.58%) [7].

Kajian penelitian yang dilakukan oleh Windari Oktapia S. dkk. Algoritma

logistic regression, dan *random forest* digunakan untuk melakukan klasifikasi emosi tweet. Dataset yang digunakan adalah Indonesian-Twitter-Emotion-Dataset. Penelitian ini menggunakan teknik SMOTE untuk mengatasi permasalahan *imbalance* data. Hasil akurasi yang didapatkan adalah algoritma *logistic regression* memperoleh nilai akurasi sebesar 78.22%, dan *random forest* memperoleh nilai akurasi sebesar 72.41%. Dapat disimpulkan bahwa algoritma *logistic regression* memperoleh akurasi klasifikasi yang lebih tinggi dari *random forest* [8].

Penelitian lain melakukan perbandingan algoritma *logistic regression* dan *naive bayes* menghasilkan algoritma *logistic regression* memiliki nilai akurasi sebesar 84.62%. Sedangkan, hasil klasifikasi algoritma *naive bayes* mendapatkan nilai akurasi sebesar 83.71% [9].

Penelitian yang dilakukan terinspirasi dari penelitian sebelumnya yang dilakukan oleh Maria I. dkk. Penelitian tersebut menggunakan metode *feature selection infomation gain* dan nilai *threshold* menggunakan nilai *standard deviation*. Hasil dalam penelitian tersebut adalah penggunaan metode *information gain* dengan *threshold* nilai *standard deviation* dan menggunakan dataset asli mendapatkan nilai rata-rata akurasi yang lebih unggul jika dibandingkan dengan metode *correlation-based feature selection* (CBFS) [6].

Selain itu, terdapat contoh penelitian yang dilakukan oleh Tsai dan Sung, yang menggunakan nilai *threshold* dari perhitungan statistik yang menghitung rata-rata setiap frekuensi atau nilai *mean* [10]. Dalam statistika, terdapat beberapa perhitungan selain nilai mean. Sebagai contoh, dalam statistika deskriptif, terdapat perhitungan *standard deviation*, modus, median, *mean*, dan lainnya [11].

Maka dari itu, penelitian ini menggunakan *information gain* untuk *feature selection* dengan *threshold* yang dihitung berdasarkan statistik nilai median. Dataset yang digunakan berasal dari UCI Machine Learning Repository dan Open ML dengan jumlah *feature* lebih dari 50. Dataset ini memiliki kategori *multivariate, sequential*, dan *univariate*. Penelitian ini melibatkan uji coba klasifikasi dari hasil *feature selection* menggunakan metode *information gain* dan nilai *threshold* median dievaluasi menggunakan algoritma *logistic regression*.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah, dapat dirumuskan dua rumusan masalah yaitu:

1. Bagaimana implementasi *information gain* untuk *feature selection* dengan

threshold nilai median?

2. Berapa nilai akurasi, presisi, *recall*,dan F1 *Score* dari implementasi *information gain* untuk *feature selection* dengan *threshold* nilai median?

1.3 Batasan Permasalahan

Pada penelitian ini, terdapat beberapa batasan yang perlu diperhatikan, yaitu:

1. Penelitian ini menggunakan 8 dataset yang memiliki kategori dataset berbeda, yakni *multivariate, univariate*, dan *sequential*. Dataset yang digunakan bersumber dari situs web UCI dan Open ML. Nama-nama dataset yang digunakan adalah AP_Breast_Omentum, Musk (Version 2), Internet Advertisements, Bioresponse, Arcene, AP_Colon_Kidney, Hill Valley, dan Nomao.
2. Setiap dataset yang digunakan memiliki lebih dari 50 fitur yang dieksplorasi dalam penelitian ini.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dijabarkan sebelumnya, maka didapatkan tujuan penelitian sebagai berikut:

1. Mengimplementasi *information gain* untuk *feature selection* dengan *threshold* nilai median.
2. Mengukur nilai akurasi, presisi, *recall*,dan F1 *Score* dari implementasi *information gain* untuk *feature selection* dengan *threshold* nilai median.

1.5 Manfaat Penelitian

Berdasarkan tujuan penelitian yang telah dijabarkan sebelumnya, maka manfaat penelitian yang didapat yaitu:

1. Memberikan sebuah informasi tentang topik *feature selection* dengan menggunakan *information gain* dan nilai *threshold* median.

2. Mengetahui performa dari hasil akurasi, presisi, *recall*,dan F1 *Score* setelah dilakukan *feature selection* yang digunakan sebagai acuan penelitian kedepannya.

1.6 Sistematika Penulisan

Berisikan uraian singkat mengenai struktur isi penulisan laporan penelitian, dimulai dari Pendahuluan hingga Simpulan dan Saran.

Sistematika penulisan laporan adalah sebagai berikut:

- Bab 1 PENDAHULUAN
  Pada bab 1 diuraikan latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

- Bab 2 LANDASAN TEORI
  Pada bab 2 menjelaskan teori-teori yang digunakan pada penelitian ini. Teori-teori yang dicantumkan meliputi *feature selection, entropy, information gain, threshold,* median.

- Bab 3 METODOLOGI PENELITIAN
  Pada bab 3 menjelaskan metodologi penelitian secara urut. Terdapat *flowchart* dari perancangan sistem.

- Bab 4 HASIL DAN DISKUSI
  Pada bab 4 menjelaskan implementasi sistem, dan hasil implementasi dari penelitian ini.

- Bab 5 KESIMPULAN DAN SARAN
  Pada bab 5 diuraikan kesimpulan dan saran dari penelitian yang telah dilakukan.