BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Big data merupakan salah satu peluang untuk digunakan sebagai *data analytic* pada suatu perusahaan. *Data analytic* menurut Andry adalah proses dalam mencari pola tersembunyi dengan korelasi yang tidak diketahui, dan informasi penting lain untuk membuat keputusan bisnis lebih informatif. Data-data yang dianalisis haruslah diukur dan dievaluasi oleh perusahaan, bukan hanya memberikan ide-ide dan tanggapan saja [1].

Perkembangan lingkup pekerjaan pada bagian arsip data perusahaan yang mengelola besarnya data yang berfokus pada situasi dan tugas baru sehingga perlu beradaptasi dengan perkembangan saat ini. Dalam proses perkembangan tersebut muncul adanya masalah dan perubahan baru sehingga meningkatkan respon dan tanggung jawab serta risiko keamanan pekerjaan dalam pencegahan dan pengedalian untuk mempromosikan tingkat spesialisasi pemrosesan file untuk mempercepat pengembangan arsip perusahaan dalam berkontribusi secara positif dan efektif [2].

Pemrosesan file pada era big data saat ini melibatkan penggabungan informasi dari berbagai sumber dengan representasi yang berbeda karena keragaman dan kualitas data sangat bervariasi dari satu sumber ke sumber lainnya, bahkan dalam bidang yang sama [3].

Dengan masalah dari segi pemrosesan file agar dapat mempercepat dan dapat berkontribusi secara efektif maka dibuatlah dengan kecerdasan buatan. Kecerdasan buatan adalah pembelajaran yang menghasilkan perhitungan yang memungkinkan untuk dapat dipahami, dinalar, dan bagaimana cara untuk bertindak. Oleh karena kecerdasan buatan menghasilkan perhitungan yang dapat dinalar, dipahami dan bertindak, maka dapat dikatakan berbeda dengan ilmu komputer, dan ilmu psikologi [4].

Kecerdasan buatan memiliki berbagai macam cabang, salah satunya machine learning. *Machine learning* pada penelitian ini berfokus pada *Supervised Classification* atau klasifikasi dimana output nya berupa rata-rata label kelas, dan nilai rata-rata evaluasi berupa akurasi, presisi, recall, Fmeasure, dan lain sebagainya [5]. Dengan diterapkan menggunakan *machine learning*, maka mampu menggeser yang awalnya *big data informatics* menjadi dunia kecerdasan [6].

Big data saat ini dengan memiliki sejumlah fitur data yang banyak yang mampu memengaruhi efektivitas prediksi, oleh karena itu fitur perlu dikurangi atau diseleksi menggunakan feature selection. Cara kerja *feature selection* sendiri ialah mengevaluasi keseluruhan atribut atau fitur dengan kriteria tertentu untuk menentukan pentingnya fitur [7].

Inti pokok bahasan penelitian ini tentang perbandingan penggunaan metode fitur seleksi *Wrapper*, yaitu *Recursive Feature Elimination* (RFE) dengan *Genetic Algorithm* (GA). Untuk *feature selection* (FS) diklasifikasi menjadi tiga jenis, yaitu "*filter, wrapper, embedded*". Pada Metode *filter* memiliki dua fase dalam proses pemilihan fitur. Pada fase pertama, yaitu menghitung atau mengukur skor kepentingan fitur berdasarkan kriteria tertentu. Kemudian pada fase kedua, dari hasil fase pertama yang memiliki skor fitur rendah akan dihilangkan. Metode *wrapper*, proses pertama dengan cara membuat model pembelajaran menggunakan sebuah subset dari fitur dan dengan berulang kali (*backward* atau *foreward*) melatih model prediksi menggunakan fitur ini. Berdasar dari hasil model tadi, maka yang tidak relevan akan dihapus. Metode *embedded* adalah metode yang proses nya menggabungkan kelebihan dari metode *wrapper* dan *filter*, dimana cara kerja nya dengan mencari hubungan ketergantungan antar fitur dan juga melatih model yang biaya nya lebih rendah dari metode *wrapper* itu sendiri [7]. Metode *wrapper* digunakan untuk menghasilkan lebih banyak solusi berbasis model [8].

Salah satu model *wrapper* yaitu Metode RFE [9]. RFE sendiri digunakan untuk meningkatkan kinerja klasifikasi dengan mengoptimalkan subset fitur [10] dengan memberikan bobot pada tiap fitur untuk menentukan pemeringkatan berdasarkan kepentingan fitur [11] dan tentu juga berdasarkan beberapa metode pembelajaran mesin tertentu yang dipilih [10].

Sejak diterapkannya pertama kali pada tahun 1989, algoritma Genetic Algorithm (GA) berperan sebagai algoritma pencarian. Kemudian algoritma ini dalam tiga tahun terakhir membuat perkembangan fantastis pada penelitian tentang Feature Selection (FS), dengan kemampuan pencarian global, mampu mengoptimalkan metaheuristik berbasis populasi [12]. Menurut Validasi eksperimental juga mengungkapkan bahwa salah satu perkembangan fantastis tersebut terjadi saat penggunaan metode tertentu dengan pengoptimalan menggunakan Genetic Algorithm (GA) memiliki akurasi prediksi sebesar 4,17%-15,14% dibandingkan dengan pemilihan fitur lainnya [13].

Oleh karena itu, berdasarkan penjelasan diatas digunakan fitur seleksi RFE dan GA untuk menyelesaikan masalah data berdimensi tinggi pada big data dengan menggunakan enam metode machine learning yaitu Logistic regression, KNN, SVM, Random Forest, AdaBoost, dan Naïve bayes. Kemudian dilakukan perbandingan hasil proses klasifikasi antara fitur seleksi menggunakan RFE dan GA dengan parameter yang digunakan sebagai kriteria penilaian perbandingan terbaik dalam masalah data berdimensi tinggi yaitu akurasi, presisi, recall, dan F-score. Kemudian hasil model klasifikasi yang diperoleh akan dibandingkan dengan input peneliti yang nantinya akan memunculkan prediksi *class* sesuai yang ada pada dataset.

1.2 Rumusan Masalah

Berdasar latar belakang masalah seperti pada bagian 1.1, berikut beberapa rumusan yang didapat.

a. Berapa hasil evaluasi akurasi, presisi, recall dan F-1 score pada proses klasifikasi seluruh dataset menggunakan fitur seleksi RFE dan GA?
b. Bagaimana hasil prediksi yang dihasilkan saat dibandingkan antara menggunakan model klasifikasi dari fitur seleksi RFE maupun GA dengan input peneliti?

1.3 Batasan Masalah

Agar masalah tidak meluas dan menyimpang, maka penulis merumuskan batasan masalah sebagai berikut.

a. Fitur seleksi yang digunakan pada penelitian ini adalah Recursive Feature Elimination (RFE) dan Genetic Algorithm (GA)
b. Algoritma yang digunakan pada penelitian ini adalah Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, AdaBoost, dan Naïve Bayes.
c. Data yang digunakan adalah dataset Wisconsin Diagnostic Breast Center (WDBC) yang membahas tentang diagnosis payudara di wilayah wisconsin. Dataset ini berasal dari UC Irvine (UCI) Machine Learning Repository. Dataset ini memiliki total data sebanyak 569 dan total 30 feature dengan class yang berisi Benign (B) dan Malignant (M).
d. Framework yang digunakan menggunakan streamlit. Streamlit digunakan untuk membangun *user interface* interaktif untuk aplikasi data science.

1.4 Tujuan Penelitian

Terdapat beberapa tujuan dari penelitian ini yaitu :

a. Penelitian ini ditujukan untuk mengetahui hasil evaluasi yang dihasilkan dari fitur seleksi terbaik yang terpilih berdasarkan test size 0.1, 0.2, dan 0.3.
b. Penelitian ini ditujukan untuk mengetahui prediksi terbaik yang diperoleh dari model klasifikasi yang telah diuji antara RFE dan GA.

1.5 Manfaat Penelitian

Terdapat manfaat dari penelitian ini yaitu :

a. Website pada penelitian ini digunakan untuk memprediksi pada data berdimensi tinggi dengan tipe file adalah csv.

1.6 Sistematika Penulisan

Sistematika penulisan dalam Tugas Akhir disusun dalam lima bab yaitu :

BAB I PENDAHULUAN, berisi latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA, berisi tinjauan pustaka, keaslian penelitian dan dasar teori yang digunakan.

BAB III METODE PENELITIAN, berisi tentang objek penelitian, alur penelitian, dan alat dan bahan.

BAB IV HASIL DAN PEMBAHASAN, berisi hasil dan pembahasan hasil penelitian yang dilakukan penulis.

BAB V PENUTUP, berisi kesimpulan dan saran yang dapat peneliti rangkum selama proses penelitian.