BAB I
PENDAHULUAN

1.1 Latar Belakang

Proses kehidupan masa ini sangat dipenuhi oleh beragam data yang dibuat dan bertumbuh seiring dengan kompleksnya kehidupan manusia. Menurut Jhon Naisbitt tahun 1988, kita telah tenggelam didalam banyaknya data namun miskin terhadap pengetahuan. Dari permasalahan ini muncul bidang ilmu baru dalam mengolah data, yang disebut data mining. *Data mining* merupakan suatu proses yang memanfaatkan metode statistik, komputasi matematis, kecerdasan buatan, serta pembelajaran mesin dalam rangka mengekstraksi dan mengidentifikasi pengetahuan dan informasi bernilai yang tersembunyi dalam beragam basis data berskala besar (Turban, 2005). Beberapa teknik telah dihasilkan untuk mengolah data dengan tujuan menerjemahkannya menjadi pengetahuan yang dapat diselidiki lebih lanjut. Salah satu metode yang relevan adalah pendekatan klasifikasi C4.5. Studi sebelumnya yang mengfokuskan pada klasifikasi data mining dengan menggunakan metode algoritma C4.5, seperti yang dilakukan oleh Laily Hermawanti (2012) dalam kajiannya berjudul "Penerapan Algoritma Klasifikasi C4.5 untuk Diagnosis Penyakit Kanker Payudara," mencatat tingkat akurasi sebesar 94,56% dan nilai *AUC* senilai 0,941. Hasil tersebut mengindikasikan tingkat akurasi yang tinggi. Studi lain yang relevan dilaksanakan oleh Siti Masripah (2015) dalam kajiannya yang berjudul "Evaluasi Penentuan Kelayakan Pemberian Kredit Koperasi Syariah Mempergunakan Algoritma Klasifikasi C4.5" mencapai tingkat akurasi senilai 88% dan nilai AUC senilai 0,898, mengindikasikan tingkat klasifikasi yang baik.

Namun pada kenyataanya Algoritma C4.5 memiliki beberapa kelemahan yaitu dapat menyebabkan terjadinya *overlap* ketika kelas atau kriteria yang digunakan jumlahnya terlalu banyak, kesulitan untuk merancang pohon keputusan yang optimal, dan hasil kualitas keputusan sangat bergantung pada desain pohon tersebut. Untuk mengatasi kelemahan algoritma C4.5 agar dapat mendesain pohon keputusan yang optimal maka untuk penelitian ini akan dioptimasi menggunakan

AdaBoost. AdaBoost merupakan teknik metode boosting yang mengkategorikan data baru dengan memanfaatkan dataset yang telah diberi bobot, dimana bobot ini berkisar antara 1 hingga 4 berdasarkan hasil pelatihan klasifikasi sebelumnya. Tujuan utamanya adalah mengurangi kesalahan dengan menyesuaikan bobot data secara iteratif. AdaBoost telah dilakukan oleh (Rohman, Suhartono, & Supriyanto, 2017) dengan mengangkat kasus prediksi penyakit jantung menggunakan algoritma C4.5 dengan AdaBoost dan juga melakukan komparasi menggunakan teknik *bagging*. Hasil yang diperoleh dari penelitian tersebut untuk prediksi penyakit jantung menggunakan C4.5 dan AdaBoost adalah 92,24% dan untuk C4.5 dan bagging adalah 91,89%. Dari penelitian yang telah diuraikan, hal tersebut belum bisa membuktikan bahwa metode *Boosting* AdaBoost sangat efektif, berdasarkan studi kasus yang dilakukan (Amalia, 2007) pada pola data tertentu AdaBoost tidak terlalu berpengaruh pada performasi algoritma C4.5.

Berikutnya kajian oleh (Debiyanti et al., 2020). Tujuan dari kajian ini ialah untuk mengantisipasi kondisi perusahaan sehubungan dengan kemampuannya dalam mengelola serta menjaga stabilitas kinerja keuangan. Kajian ini menerapkan pendekatan *Decision Tree* dengan metode C4.5 yang diperkuat oleh *Adaptive Boosting*. Awalnya, kajian ini memanfaatkan 755 catatan data, namun setelah melalui proses KDD, jumlah catatan data yang tersedia berkurang menjadi 746. Hasil pengujian yang telah dilakukan dengan menggunakan perbandingan data latih dan data uji sebesar 90%:10% menunjukkan bahwasanya akurasi Algoritma C4.5 mencapai 72,97%. Namun, setelah ditingkatkan melalui *Adaboost,* akurasinya mengalami peningkatan menjadi 86,49%. Maka, bisa ditarik simpulan bahwasanya penggabungan Algoritma C4.5 dengan *Adaboost* adalah pendekatan yang efektif dalam memprediksi risiko keuangan perusahaan. Berlandaskan pada penyelesaian masalah yang ada, algoritma C4.5 berbasis *Adaptive Boosting* (AdaBoost) tidak begitu signifikan dalam melakukan peningkatan performasi algoritma C4.5. Sehingga diperlukan suatu penelitian yang menerapkan parameter dan perlakuan yang sama pada dataset yang berbeda untuk mengukur seberapa besar keefektifan AdaBoost untuk menumbuhkan performa algoritma C4.5.

Untuk mengetahui tingkat *performance* tersebut, diperlukan studi kasus dari data yang berbeda, ukuran ataupun karakteristik data yang berbeda, dan juga dengan atribut serta beberapa hasil data yang ada pada studi kasus yang berbeda. Selain itu merujuk pada performasi sebuah algoritma, jumlah variabel atau fitur sangat mempengaruhi efektifitas algoritma dalam bekerja. Hal ini berkaitan pada jumlah fitur serta biaya komputasi yang dibutuhkan dalam proses penyelesaian masalah(Kotu & Deshpande, n.d.). Sehingga pada penelitian ini peneliti juga melakukan proses selektif fitur untuk meningkat akurasi serta performasi penelitian. Selektif fitur merupakan proses dalam analisis data dan pemodelan statistik di mana subset terbaik dari fitur atau atribut dari kumpulan data yang ada dipilih untuk digunakan dalam pembangunan model. Seleksi fitur bertujuan untuk mengidentifikasi dan mempertahankan fitur-fitur yang paling relevan dan informatif, sementara menghilangkan fitur-fitur yang *redundancy*, tidak relevan, atau memiliki dampak minimal terhadap prediksi atau variabel target. Seleksi fitur dilakukan dengan menggunakan berbagai teknik dan metode, dalam penelitian ini, peneliti menggunakan Teknik Filter yaitu proses PCA, *Information Gain* dan Proses Chi Square dan Teknik Wrapper yaitu proses *forward* dan proses *backward*. Untuk kasus pada penelitian kali ini, penulis menggunkan 4 (Empat) datashet yang berbeda diantaranya ialah *Airline Passenger Satisfaction, IRIS, TIC TAC TU, Water_Potability*. Dimana semua dataset tersebut diambil dari Kiggle. Dengan memberikan perlakuan yang sama pada setiap dataset membuat data tersebut lebih terarah dan mudah untuk di analisis serta mempermudah untuk mengetahui seberapa besar efektifitas dari AdaBoost dalam meningkatkan performasi algoritma C4.5 dalam menangani suatu data.

Temuan yang diperoleh pada kajian ini berupa model evaluasi dari tahapan masing-masing sampel kasus klasifikasi sebelum dan sesudah menggunakan algoritma C4.5 berbasis AdaBoost pada setiap dataset yang berbeda, dengan perlakuan yang sama sehingga dapat diketahui secara detail seberapa besar tingkat keefektifan AdaBoost dalam meningkatkan performasi algoritma C4.5. Sehingga didapatkan tahapan dalam *preprocessing* yaitu menerapkan slektif fitur yang sesuai pada kriteria setiap dataset.

1.2 Identifikasi Masalah

Identifikasi permasalahan pada kajian Perbaikan *Performance* Algoritma C4.5 Adaboost Berbasis Optimasi Proses Selektif Fitur yang dapat diajukan adalah sebagai berikut.

1. Algoritma C4.5 memiliki suatu kelemahan dimana semua data *training* yang ada harus disimpan pada penyimpanan dan dalam waktu yang bersamaan. Selain itu masalah *overfitting* diakibatkan oleh terjadinya misklasifikasi sebab *noisy* data, tidak seimbangnya data mengakibatkan tingkat akurasi rendah pada algoritma C4.5 dalam hal pengklasifikasian data. Sehingga algortima C4.5 perlu di optimalkan dengan cara pemberian bobot.
2. Banyak penelitian yang hanya berfokus untuk mengimplementasikan algoritma C4.5 berbasis AdaBoost pada satu objek saja, sehingga belum diketahui secara pasti apakah dengan memberikan bobot berupa AdaBoost dapat meningkatkan performasi algoritma C4.5 secara signifikan atau tidak.

1.3 Rumusan Masalah

Berlandaskan pada latar belakang dalam Perbaikan *Performance* Algoritma C4.5 Adaboost Berbasis Optimasi Proses Selektif Fitur diatas bisa dibuat rumusan permasalahan meliputi:

1. Bagaimana efektifitas metode *Adaptive Boosting* (AdaBoost) pada algoritma C4.5 untuk mengklasifikasi suatu data?
2. Bagaimana efektifitas penerapan selektif fitur pada Algoritma C4.5 dan C4.5 berbasis AdaBoost?
3. Faktor apa saja yang mempengaruhi tingkat akurasi yang dihasilkan dari algoritma C4.5 ?

1.4 Batasan Masalah

Supaya interpretasi dari permasalahan tidak melebar dari tujuan penulisan, maka dibuat sejumlah batasan yang perlu dibuat dalam melakukan penelitian Perbaikan *Performance* Algoritma C4.5 Adaboost Berbasis Optimasi Proses Selektif Fitur:

1. Penelitian ini berfokus pada seberapa besar performasi dan efektifitas penerapan *Adaptive Boosting* (AdaBoost) pada algoritma C4.5 dalam meningkatkan performasi algoritma C4.5 pada suatu klasifikasi data.
2. Terdapat Empat dataset yang berbeda dari segi karakteristik dan tipe data yang berbeda guna menguji seberapa efektif penerapan algoritma C4.5 berbasis AdaBoost. Dataset yang digunakan berjumlah 4 buah dataset, diantaranya ialah *Airline Passenger Satisfaction, IRIS, TIC TAC TU* dan *Water_Potability*.
3. Pengujian dilaksanakan melalui *confusion matrix* dan AUC
4. Pengujian dilaksanakan melalui Aplikasi Rapidminer.

1.5 Tujuan Penelitian

Adapun tujuan dari penelitian Perbaikan *Performance* Algoritma C4.5 Adaboost Berbasis Optimasi Proses Selektif Fitur meliputi

1. Untuk mengetahui tingkat efektifitas dari algoritma C4.5 berbasis *AdaBoost* dalam meningkatkan performasi algoritma C4.5 pada suatu klasifikasi data mining.
2. Untuk mengetahui tingkat efektifitas penerapan selektif fitur pada Algoritma C4.5 dan C4.5 berbasis AdaBoost.
3. Untuk mengetahui perbandingan akurasi teknik *data mining* Algoritma C4.5 berbasis AdaBoost pada setiap Dataset.

1.6 Manfaat Penelitian

Adapun manfaat yang dioperoleh bagi peneliti dari penelitian Perbaikan *Performance* Algoritma C4.5 Adaboost Berbasis Optimasi Proses Selektif Fitur adalah

a. Manfaat Teoretis
   1. Untuk penulis, harapannya dapat menambah wawasan memahami serta bisa mengimplementasikannya pada kegiatan pembelajaran yang tengah dilaksanakan.
   2. Untuk kajian yang serupa, kajian ini harapannya bisa dijadikan referensi untuk membantu penerapan kajian terkait.

b. Manfaat Praktis

1. Untuk penulis bisa mengaplikasikan ilmu pengetahuan yang sudah diperoleh ketika berkuliah pada Analisis penerapan Algoritma *C 4.5* Berbasis AdaBoost pada beberapa dataset.