BAB I

PENDAHULUAN

1.1 Latar Belakang

Film merupakan sebuah media yang dapat berperan banyak dalam kehidupan sehari-hari, sebagai contoh sebuah film yang dilihat oleh seorang seniman bisa saja memberikan sebuah inspirasi untuk berkarya lebih baik lagi, begitupun dengan film tentang kehidupan bisa mengajarkan seseorang anak untuk melihat betapa kerasnya perjuangan orang tua untuk menghidupi anaknya sampai dia dewasa. Akan tetapi, film juga mempunyai sisi negatif jika orang yang melihat film tersebut tidak mengikuti anjuran yang sudah tertera seperti batasan umur untuk menonton sebuah film.

Banyaknya film bajakan yang tersebar luas di internet dan mudahnya kita dalam mengaksesnya menimbulkan berbagai permasalahan, salah satunya adalah apakah film yang ditonton sudah sesuai dengan batasan umur untuk menonton film tersebut. Hal ini bisa kita liat dari kehidupan sehari-hari yang terjadi di sekitar kita, contohnya adalah perilaku anak dibawah umur yang berperilaku tidak sesuai dengan umur mereka. Tentu saja hal itu menimbulkan sebuah pertanyaan apakah film yang mereka lihat menjadi salah satu penyebab mereka berperilaku tidak sesuai dengan umur mereka atau tidak. Karena itulah untuk bisa mengetahui apakah suatu film layak ditonton atau tidak berdasarkan batasan umur diperlukan sebuah data yang berisi daftar film terbaru dan diolah untuk mengetahui film-film apa saja yang boleh ditonton untuk anak dibawah umur.

Masalah utamanya adalah bagaimana cara data-data dari koleksi yang ada diolah untuk dikelompokkan sesuai dengan yang sudah ditentukan agar tidak menimbulkan kesalahan dalam memilih film yang layak untuk ditonton karena ketidaktahuan jenis-jenis tontonan yang diperbolehkan kepada anak dibawah umur. Data yang digunakan dalam penelitian ini merupakan data digital koleksi film yang terdapat di internet yang berasal dari sumber terpercaya. Sampel yang digunakan adalah sampal dari situs database scientist ternama bernama kaggle yang sudah banyak digunakan oleh para ahli dalam melakukan berbagai penelitian.

Untuk mengatasi masalah tersebut solusi yang diterapkan adalah melakukan

penerapan metode sistem *data mining*. *Data mining* merupakan suatu proses dimana data dalam jumlah besar dianalisis dan dieksplorasi dengan tujuan untuk menemukan pola atau aturan dengan cara otomatis maupun semiotomatis[1]. Metode sistem *data mining* sendiri teradapat beberapa jenis seperti klasifikasi, assosiasi, klasterisasi, regresi, dan masih banyak lagi[1]. Akan tetapi, fokus metode yang digunakan dalam penelitian ini adalah metode *data mining classification* dengan algoritma C5.0 dikarenakan memudahkan untuk melakukan penggalian informasi dimana hasil dari data yang diklasifikasikan dalam bentuk pohon keputusan maupun aturan *if then* dapat diperoleh dengan baik dan jelas[2].

Algoritma C5.0 merupakan algoritma yang dapat menangani atribut kontinyu dan diskrit yang disempurnakan dari algoritma sebelumnya yaitu algoritma ID3 dan C4.5 yang berbasis *decision tree* yang dibentuk pada tahun 1987 oleh Ross Quinlan. Kesalahan yang ditimbulkan dalam pengambilan keputusan pada algoritma ini lebih diminimalkan karena pemilihan atribut dalam algoritma ini akan diproses menggunakan information *gain*[2].

Beberapa penelitian sudah dilakukan tentang algoritma C5.0 seperti penelitian tentang perbandingan performansi algoritma *Decision Tree* C5.0, CART, dan CHAID[3], Implementasi Algoritma C5.0 pada Penilaian Kinerja Pegawai Negeri Sipil[4], *The Decision Tree C5.0 Classification Algorithm for Predicting Student Academic Performance* oleh N. Benedeiktus dan R. Oetama[5], Penerapan Algoritma C5.0 Pada Sistem Pendukung Keputusan Kelayakan Penerimaan Beras Masyarakat Miskin[6], Implementasi Algoritma C5.0 Pada Kelulusan Peserta Ujian Kemahiran Berbahasa Indonesia (Ukbi) Pada Balai Bahasa Sumatera Utara[7], Implementasi Algoritma C5 . 0 Dalam Klasifikasi Pendapatan Masyarakat[1], dimana didapat bahwa tingkat akurasi algoritma C5.0 dalam melakukan *credit scoring* lebih baik dibandingkan dengan algoritma lainnya[3], [4].

Berdasarkan beberapa penelitian yang sudah dilakukan sebelumnya ditemukan beberapa kelebihan dan kekurangan dari algoritma C5.0 seperti algoritma C5.0 memiliki kelebihan untuk menangani *missing value* dan data dengan jumlah yang besar sehingga menyebabkan algoritma C5.0 dapat melakukan *training* data dalam waktu

yang cepat untuk digunakan dalam testing data. Selain itu, hasil dari pohon keputusan menggunakan algoritma C5.0 dapat dipangkas atau terdapat *pruning* (pemangkasan). *Pruning* dilakukan ketika *tree* yang dibangun memiliki *overfitting. Overfitting* sendiri merupakan sebuah kondisi dimana pada saat pembangunan *tree* terdapat noise sehingga menyebabkan *tree* yang dibangun menjadi lebih besar. Walaupun memiliki keunggulan dalam kecepatan untuk *training* data algoritma C5.0 tidak memiliki masalah dalam hal akurasi karena memiliki metode *boosting* yang dapat digunakan untuk meningkatkan tingkat akurasi. Namun dibalik kelebihan algoritma C5.0 dalam mengolah data, algoritma C5.0 sangat bergantung pada informasi dimana perubahan kecil dalam inputan data dapat menyebabkan perubahan yang besar pada pohon keputusan, dan juga jika data yang digunakan memiliki varian yang sangat kecil maka model prediksi dari algoritma C5.0 akan menjadi tidak stabil[5], [8].

Dari beberapa hasil penelitian yang sudah pernah dilakukan tersebut, maka penulis tertarik untuk membuat penelitian ilmiah dengan mengusulkan metode algoritma C5.0 sebagai fokus utama penelitian dengan judul "Klasifikasi Data FILM untuk Menentukan Rekomendasi FILM untuk anak menggunakan Algoritma C5.0" dengan tujuan utama dari penelitian ini adalah bagaimana penerapan metode *data mining* dalam membantu proses untuk mengkategorikan dan merekomendasikan film agar mempermudah para orang tua dalam memilih film yang ditonton dan dinikmati oleh anak sesuai dengan anjuran yang ada.

1.2 Rumusan Masalah

- Apakah banyaknya varian data pada dataset akan mempengaruhi performa dari model klasifikasi yang dibuat menggunakan algoritma C5.0?
- Bagaimana hasil performa akurasi dari model klasifikasi algoritma C5.0 menggunakan pruning?

1.3 Tujuan Penelitian

Tujuan penelitian yang ingin dicapai dalam penelitian ini adalah:

- Untuk mengetahui performa terbaik model klasifikasi algoritma C5.0 berdasarkan evaluasi metriks

- Untuk mengetahui performa terbaik model klasifikasi algoritma C5.0 dengan menggunakan Pruning

1.4 Batasan Masalah

Untuk menghindari adanya pelebaran pokok masalah dalam penelitian, pembatasan suatu masalah dibutuhkan agar penelitian yang dilakukan menjadi lebih terarah. Adapun beberapa batasan masalah dalam penelitian ini dapat dilihat pada poin-poin berikut:

- Luas lingkup penelitian berfokus kepada penyajian informasi tentang proses pengklasifikasian data film menggunakan model klasifikasi algoritma C5.0
- Luas lingkup penelitian berfokus pada performa model klasifikasi algoritma C5.0 yang digunakan. Hasil klasifikasi individual tidak ditampilkan untuk menjaga kesederhanaan dan fokus pada performa keseluruhan model.