BAB III
METODOLOGI PENELITIAN

3.1. Gambaran Umum Objek Penelitian

Objek yang diteliti pada penelitian ini merupakan *rating* film-film Indonesia. *Rating* adalah elemen penting untuk menentukan kualitas dari sebuah film. *Rating* merupakan kesimpulan yang menggambarkan kualitas dari elemen-elemen yang ada dalam sebuah film [4]. *Rating* dari sebuah film bisa didapatkan berdasarkan *feedback* dari masyarakat yang telah menonton film tersebut [7]. Diperlukannya prediksi *rating* film ini adalah untuk membantu perkembangan dunia perfilman di Indonesia. Bagi masyarakat, prediksi *rating* ini dapat berguna untuk referensi untuk menonton film karena biasanya film dengan *rating* yang tinggi akan lebih diminati. Bagi pembuat film, prediksi *rating* ini akan berguna untuk pengambilan keputusan yang nantinya dapat menghasilkan manfaat untuk meningkatkan kualitas film yang dibuat dan mengetahui minat masyarakat.

3.1 Metode Penelitian

3.1.1 Data Collection

Data yang digunakan berasal dari dataset *rating* film Indonesia yang berasal dari website Kaggle.com yang dikumpulkan oleh Dionisius Darryl Hermansyah dari Institut Teknologi Bandung dengan judul IMDb Indonesian Movies. Data-data berasal dari website IMDb.com (Internet Movie Database) menggunakan tools IMDb-scraper yang kemudian diubah dan dibersihkan ke dalam bentuk csv.

Dataset *rating* film Indonesia ini berisikan variable-variabel berbentuk numeric dan nominal yang berpengaruh terhadap *rating* yang akan didapatkan dari sebuah film. Variabel-variabel yang ada pada dataset *rating* film Indonesia adalah *title, year, description, genre, rating,, votes, languages, directors, actors, dan runtime, user rating.* Variabel-variabel ini juga terbagi dalam dua

tipe data yaitu Nominal dan *Numeric*. *Numeric* merupakan data yang berbentuk angka sedangkan nominal merupakan data yang bersifat kategorik [16].

Tabel 3. 1 Tipe Data

No.	Atribut	Keterangan	Tipe
1	*title*	Judul film	Nominal
2	*year*	Tahun rilis	Numeric
3	*description*	Deskripsi film	Nominal
4	*genre*	*Genre* film	Nominal
5	*MPAA_rating*	Sertifikasi *rating* umur	Nominal
6	*votes*	Jumlah vote	Numeric
7	*languages*	Bahasa film	Nominal
8	*directors*	Sutradara	Nominal
9	*actors*	Actor / aktris	Nominal
10	*runtime*	Panjang waktu film	Nominal
11	*user rating*	IMDb *user rating*	Numeric

3.1.2 Variabel Independen

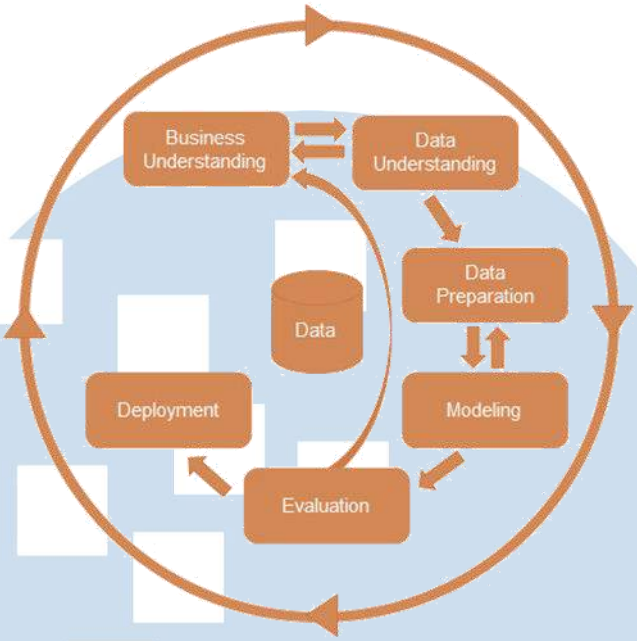
Variabel independen merupakan variabel yang nilainya mempengaruhi variabel dependen atau disebut juga sebagai variabel bebas karena nilainya tidak bergantung dengan variabel lainnya. Variabel independen pada *dataset* yang digunakan disini adalah *title, year, description, genre, MPAA_rating, votes, languages, directors, actors, dan runtime.*

3.1.3 Variabel Dependen

Variabel dependen merupakan variabel yang nilainya dipengaruhi oleh variabel independen atau bisa juga disebut variabel yang memiliki nilai terikat. Variabel dependen pada dataset yang digunakan disini adalah *users_rating*. Variabel *users_rating* berbentuk angka dengan skala 1 hingga 10 yang diberikan oleh para penonton. Nilai 1 menandakan film tersebut merupakan film yang buruk dan merupakan salah 1 film terburuk yang pernah ditonton dan nilai 10 menandakan film tersebut adalah film yang baik atau bagus.

3.2 Alur Penelitian

Penelitian ini akan menggunakan metode dari *data mining* yaitu Cross Industry Process for *Data mining* atau yang biasanya lebih dikenal dengan singkatan CRISP-DM. Metode CRISP-DM dipilih karena merupakan metode populer yang dapat meningkatkan kesuksesan sebuah proyek *data mining*. Metode CRISP-DM berbentuk siklus yang terdiri atas enam tahap yang akan membantu implementasi dari *data mining* untuk diaplikasikan dalam lingkungan yang nyata seperti membantu untuk pengambilan keputusan bisnis [11]. Flowchart dari metode CRISP-DM dapat dilihat pada gambar 3.1 [12].



Gambar 3. 1 CRISP-DM [12]

3.2.1 Business Understanding

Tahap pertama yang dilakukan dalam metode CRISP-DM adalah Business Understanding. Business understanding dilakukan untuk merencanakan dan memahami tujuan yang ingin dicapai dalam penelitian [12].

3.2.2 Data Understanding

Data yang digunakan berasal dari dataset *rating* film Indonesia yang berasal dari website Kaggle.com yang dikumpulkan oleh Dionisius Darryl Hermansyah dari Institut Teknologi Bandung dengan judul IMDb Indonesian Movies. Data-data berasal dari website IMDb.com (Internet Movie Database) menggunakan tools IMDb-scraper yang kemudian diubah dan dibersihkan ke dalam bentuk csv.

Dataset ini berisikan data *rating* film Indonesia dari tahun 1926 hingga 2020 yang berjumlah sebanyak 1272 judul film didalamnya. Dataset IMDb Indonesian Movies ini juga memiliki atribut sebanyak 11, 8 atribut bertipe nominal dan 3 atribut bertipe numeric. Atribut-atribut ini nantinya akan digunakan dalam proses pembuatan model untuk prediksi *rating* film Indonesia. Beberapa missing value juga masih ditemukan dalam dataset IMDb Indonesian Movies ini. Missing value ini perlu diperbaiki di tahap data

preparation dengan melakukan preprocessing data sehingga dataset ini dapat digunakan untuk tahap permodelan nantinya.

3.2.3 Data Preparation

Tahap data preparation ini adalah untuk menyiapkan dataset sehingga dapat digunakan untuk permodelan. Pada tahap ini dilakukan preprocessing data dengan tujuan untuk memperbaiki dan menghasilkan data yang baik dan siap digunakan untuk tahap modeling selanjutnya. Data preprocessing yang dilakukan pada penelitian ini adalah melakukan filter dan data cleansing.

Filter data dilakukan untuk memilih data yang akan digunakan, dataset yang dipakai berawal dari tahun 1926 sedangkan untuk penelitian ini hanya akan digunakan data dari tahun 2011 hingga 2020. Setelah data di filter akan dilakukan data cleansing untuk memperbaiki missing value yang ada dalam atribut-atribut pada dataset. Untuk mengatasi missing value yang ada salah satu caranya adalah dengan melakukan penghapusan rows pada data yang memiliki missing value di dalamnya. Setelah data di filter dan cleansing, dataset yang baru ini sudah siap digunakan untuk tahap selanjutnya yaitu modeling atau permodelan.

3.2.4 Modeling

Tahap modeling adalah tahap dimana akan dilakukannya pengaplikasian teknik *data mining* untuk membuat sebuah model. Teknik *data mining* yang dipilih untuk penelitian ini adalah Naïve Bayes dan KNN. Tools yang dipakai untuk implementasi adalah Rapid Miner.

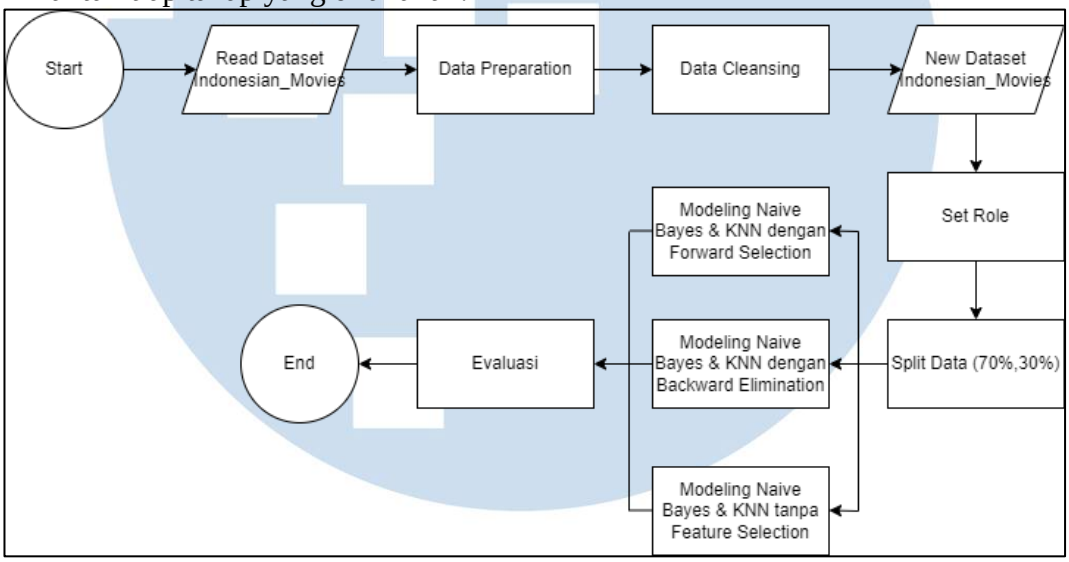
Pemilihan teknik Naïve Bayes dikarenakan Naïve Bayes merupakan algoritma yang populer untuk memprediksi *rating* dari suatu film. Berdasarkan penelitian terdahulu, Naïve Bayes digunakan sebanyak dua kali dengan hasil akurasi yaitu 53,99% [7], 55,80% [5], dan sedangkan KNN dipilih sebagai algoritma pembanding dalam penelitian ini.

Naïve Bayes merupakan algoritma yang bekerja dengan cara menghitung sekumpulan probabilitas dengan menjumlahkan menjumlahkan frekuensi dan

kombinasi nilai dari dataset yang diberikan [17]. Berbeda dengan algoritma Naïve Bayes, algoritma KNN bekerja dengan cara mencari jarak terdekat di antara data yang akan dievaluasi dengan *k-neighbor* dalam data *training* [17].

Gambar 3.2 merupakan *flowchart* atau alur dari tahap permodelan algoritma Naïve Bayes dan KNN berdasarkan [5].

Berdasarkan *flowchart modeling* diatas, berikut merupakan penjelasan untuk tiap tahap yang dilakukan:



Gambar 3. 2 Flowchart Modeling

a. Memulai dengan membaca dataset Indonesian_Movies yang masih mentah.
b. *Data preparation*, yaitu menyiapkan data dengan memilih atribut-atribut yang akan digunakan untuk *modeling*.
c. *Data cleansing* dan *handle missing value* jika ada.
d. Dataset baru yang telah dihasilkan akan siap digunakan untuk *modeling*.
e. *Set role* pada atribut yang akan dijadikan *label*.
f. Membagi data menjadi data training sebesar 70% dan data testing sebesar 30%.
g. Membuat model dengan metode *backward elimination* untuk algoritma Naïve Bayes dan KNN.
h. Evaluasi performa dari masing-masing model yang telah dibuat.

Tools yang akan digunakan pada tahap modeling ini adalah RapidMiner. Rapid Miner dipilih karena memiliki banyak kelebihan dibanding WEKA yang salah satunya dapat menggunakan operator WEKA melalui extension [14]. Rapid Miner juga dapat menangani data dengan ukuran yang besar sedangkan WEKA tidak dapat dikarenakan *memory* yang terbatas [18]. Berikut merupakan tabel perbandingan tools menurut [14] dan [18].

Tabel 3. 2 Perbandingan Tools

Tools	Kelebihan	Kekurangan
Rapid Miner	• *User-friendly GUI*. • *Integrated environment*. • Bisa menangani data dengan ukuran yang besar. • Memiliki *application wizards* yang dapat membuat model secara otomatis berdasarkan kebutuhan proyek. • Bisa menggunakan sebagian besar operator WEKA melalui *extension*.	• Dukungan untuk metode *deep learning* dan beberapa *advanced machine learning* masih terbatas.
WEKA	• *Open source*. • *User-friendly*. • Mendukung banyak prosedur evaluasi model.	• Tidak bisa menangani data yang besar dikarenakan *memory* yang terbatas. • Lambat dan *resource demanding* pada beberapa implementasi algoritma DM. • Kurangnya metode visualisasi. • Dukungan untuk *big data*, *text mining*, dan *semi-supervised learning* masih terbatas. • Tidak ada dukungan untuk *deep learning*.

3.2.5 Evaluation

Tahap Evaluation adalah tahap dimana dilakukannya evaluasi model yang telah dibuat pada tahap sebelumnya apakah sudah sesuai dengan tujuan bisnis yang ada pada tahap Business Understanding. Data testing yang berjumlah sebesar 30% dari total dataset akan digunakan untuk menguji dan akan divalidasi dengan menggunakan metode 10-*fold cross validation*.

3.2.6 Deployment

Tahap Deployment menurut [12] dapat dilakukan dengan menghasilkan laporan akhir yang telah dibuat berdasarkan *deployment plan*. Pada penelitian ini, peneliti tidak mengadakan tahap Deployment karena penelitian ini tidak diimplementasikan kepada perusahaan, namun hanya untuk keperluan studi.