

# BAB I

## PENDAHULUAN

### 1.1 LATAR BELAKANG

Dunia tengah waspada penyebaran sebuah virus yang dikenal virus corona. Penyakit yang ditimbulkan virus corona, atau dikenal dengan Covid-19, merupakan jenis baru yang ditemukan pada tahun 2019 dan belum pernah diidentifikasi menyerang manusia sebelumnya[1]. Perkara virus corona mulai muncul dan menyerang manusia pertama kali di provinsi Wuhan, China. Virus corona merupakan penyakit *zoonosis*, sehingga ada kemungkinan virus tersebut berasal dari hewan dan akhirnya menular pada manusia. Kemunculan awal tanda-tanda dari gejala yang ditunjukkan oleh virus ini umumnya diawali dengan flu, batuk, demam, sesak napas serta tidak nafsu makan dan mudah lelah. Akan tetapi gejala ini berbeda dengan *influenza*. Berdasarkan pernyataan yang dikeluarkan, virus corona mampu berkembang dengan cepat hingga menyebabkan infeksi lebih parah dan gagal organ pada manusia. Sesuai data *World Health Organization* pada tahun 2020 terkonfirmasi 6.653.469 kasus aktif dan 159.735 kasus meninggal. Virus ini telah menyebar hampir semua negara, termasuk Indonesia. Indonesia pertama kali mengkonfirmasi kasus Covid-19 ini pada Senin, 2 Maret 2020. Penularan virus secara langsung menyebar antar manusia dan menyebabkan peningkatan jumlah kasus Covid-19 meningkat secara drastis. Sejauh ini terdapat lebih dari 100 kandidat vaksin Covid-19 yang sedang dikembangkan, dan

diantaranya masih pada tahap uji coba[2]. Tetapi vaksin yang benar-benar efektif untuk mengatasi Covid-19 belum ditemukan. Berdasarkan dari permasalahan

Covid-19 tersebut, untuk meminimalisir tingkat kasus penyebarannya kasus Covid-19 masyarakat wajib melakukan pembatasan sosial (*Sosial Distancing*), membatasi aktifitas pada luar tempat tinggal, serta pemberlakuan *Lock Down* sebagai antisipasi kenaikan jumlah penderita Covid-19. Oleh karena itu dibutuhkan solusi teknologi berbasis data secara otomatis yang dapat membantu mengklasifikasikan tingkat penyebaran kasus Covid-19 di Indonesia.

*Data mining* adalah teknik yang digunakan untuk membangun model pembelajaran mesin. Pembelajaran mesin (*Machine Learning*) adalah teknik kecerdasan buatan modern yang belajar membangun model dengan menggunakan data empiris. *Data Mining* digunakan untuk menemukan pola dalam kumpulan besar data mentah. *Data mining* menerapkan teknik *Machine Learning* untuk menarik pengetahuan pada data[3].

Penelitian ini menerapkan teknik *Data mining* untuk mengklasifikasikan dataset Covid-19 menggunakan Algoritma *Naiïve Bayes*, dengan tujuan dari klasifikasi data ini adalah mengelompokkan sifat-sifat yang sama kedalam kelompok-kelompok tertentu atau kelas-kelas tertentu. Sehingga dapat mempermudah untuk membandingkan, mengelompokkan informasi yang menonjol dan menghilangkan hal-hal yang tidak perlu, menunjukkan sifat-sifat yang menonjol sehingga mudah untuk dilihat secara sekilas, dan mempermudah mengadakan perlakuan secara statistik terhadap data yang telah dikumpulkan.

Salah satu masalah pada klasifikasi adalah banyaknya atribut yang digunakan pada sebuah dataset. Atribut yang banyak membuat nilai akurasi menjadi rendah, maka atribut yang ada harus dipilih dengan algoritma yang tepat. Pemilihan Metode *Naïve Bayes* didasari beberapa alasan seperti kemudahan dalam implementasinya, dan memiliki performa cepat dalam melatih, serta dapat digunakan pada dataset yang lebih besar. Metode pengklasifikasi *Naïve Bayes* merupakan metode yang sangat sederhana dan efisien. Disamping kesederhanaannya, *Naïve Bayes* masih memiliki beberapa kekurangan yakni sangat sensitif dalam pemilihan fitur. Jumlah fitur yang terlalu banyak tidak hanya dapat meningkatkan waktu perhitungan tetapi juga dapat menurunkan akurasi klasifikasi. Tingkatan lain yang umumnya ditemukan dalam pendekatan klasifikasi sentimen adalah pemilihan fitur. Pemilihan fitur bisa membuat pengklasifikasi baik lebih efisien atau efektif dengan mengurangi jumlah data yang dianalisis. Beberapa metode pemilihan fitur yang sering digunakan antara lain *Document Frequency*, *Mutual Information*, *Information Gain*, dan *Chi-Square*.

Uysal dan Gunal [4], menyatakan bahwa :

“*Information Gain* sering lebih unggul dibanding yang lainnya. *Information Gain* mengukur berapa banyak informasi kehadiran dan ketidakhadiran dari suatu kata yang berperan untuk membuat keputusan klasifikasi yang benar dalam class apapun. *Information Gain* adalah salah satu pendekatan filter yang sukses dalam pengklasifikasian”.

Penelitian yang sudah dilakukan dalam pengklasifikasian sentimen menggunakan klasifikasi *Naïve Bayes*, diantaranya adalah penelitian oleh Alvina Felicia Watratan et al[5], mereka melakukan prediksi tentang Implementasi

Algoritma *Naïve Bayes* Untuk Memprediksi Tingkat Penyebaran Covid-19 di Indonesia menggunakan metode *Naïve Bayes* untuk menghasilkan probabilitas setiap kriteria untuk *class* yang berbeda, sehingga nilai-nilai probabilitas dari kriteria tersebut dapat dioptimalkan untuk memprediksi tingkat penyebaran Covid-19 di Indonesia berdasarkan proses klasifikasi. Berdasarkan dari penelitian sebelumnya data kasus Covid-19 Per Provinsi yang dijadikan data training diuji menggunakan metode *Naïve Bayes* memprediksi tingkat penyebaran kasus Covid-19 Per Provinsi dengan persentase keakuratan sebesar 48,4848%.

Penelitian ini ingin mengoptimasi penelitian sebelumnya dan membandingkan hasil persentase dari penelitian Felicia Watratan et al pada tahun 2020[5] yang menghasilkan persentase sebesar 48,4848% dan masih menggunakan metode yang sama seperti sebelumnya dengan metode *Naïve Bayes* serta dikombinasikan dengan seleksi fitur yaitu menggunakan *Information Gain* sebagai pemilihan fiturnya. Metode *Naïve Bayes* dan *Information Gain* sebagai pemilihan fiturnya diharapkan dapat mengoptimasi dan membandingkan hasil persentase dari penelitian sebelumnya untuk mengklasifikasikan penyebaran tingkat kasus Covid-19 di Indonesia.

Berdasarkan uraian permasalahan diatas, penelitian ini tertarik untuk melakukan penelitian dengan judul ” **PENERAPAN DATA MINING UNTUK MENGOPTIMASI KLASIFIKASI TINGKAT PENYEBARAN COVID-19 DI INDONESIA MENGGUNAKAN METODE NAÏVE BAYES DENGAN SELEKSI FITUR INFORMATION GAIN** ”.

## **1.2 RUMUSAN MASALAH**

Berdasarkan latar belakang diatas maka dapat dirumuskan permasalahan yang akan diteliti yaitu :

1. Bagaimana cara menerapkan Metode *Naïve Bayes* yang dikombinasikan dengan metode *Information Gain* untuk analisis dalam mengklasifikasikan tingkat kasus Covid-19 di Indonesia ?
2. Bagaimana tingkat akurasi pada proses klasifikasi menghitung keakurasian dengan menggunakan metode *Naïve Bayes* yang ditambahkan dengan metode *Information Gain* ?

## **1.3 BATASAN MASALAH**

Agar dapat lebih fokus dan pembahasan tidak menyimpang dari permasalahan yang ada, maka penelitian ini membatasi masalah sebagai berikut :

1. Penelitian ini hanya berfokus pada data yang diambil dari web kaggle.com.
2. Metode yang digunakan adalah *Naïve Bayes* dan dikombinasikan dengan metode *Information Gain*.
3. Alat bantu analisa menggunakan WEKA.

## **1.4 TUJUAN DAN MANFAAT PENELITIAN**

### **1.4.1 Tujuan Penelitian**

Agar dapat mengacu pada permasalahan yang ada pada penelitian ini, penelitian ini menetapkan beberapa tujuan dari penelitian ini yaitu :

1. Untuk menghasilkan data klasifikasi tingkat kasus Covid-19 di Indonesia menggunakan metode *Naïve Bayes* dan menggunakan metode *Information Gain*.
2. Untuk mengetahui keakurasian klasifikasi menggunakan metode *Naïve Bayes* yang ditambahkan dengan metode *Information Gain*.
3. Untuk memberikan informasi mengenai data tingkat kasus Covid-19 agar lebih waspada dalam penyebaran Covid-19.

#### **1.4.2 Manfaat Penelitian**

Adapun manfaat dari penelitian ini yaitu :

1. Diharapkan dengan adanya penelitian ini dapat menjadi informasi bagi pembaca untuk mengetahui tingkat penyebaran kasus Covid-19 di Indonesia.
2. Bisa menambah pengetahuan peneliti tentang bagaimana mencari informasi penting yang tersembunyi dalam suatu data menggunakan metode *Naïve Bayes* dan ditambahkan seleksi fitur dengan metode *Information Gain*.

#### **1.5 SISTEMATIKA PENULISAN**

Untuk memberikan suatu gambaran yang jelas mengenai isi penulisan tugas akhir yang akan disusun, maka dibuatlah sistematika penulisan yaitu sebagai berikut :

#### **BAB I : PENDAHULUAN**

Pendahuluan menjelaskan tentang latar belakang masalah, rumusan masalah, batasan masalah, tujuan dan manfaat penulisan, dan sistematika penulisan.

**BAB II : LANDASAN TEORI**

Landasan teori ini membahas tentang teori-teori serta pendapat dari para ahli yang berhubungan dengan permasalahan yang dianalisis dan kajian-kajian penelitian sebelumnya yang relevan dengan penelitian ini. Penulisan kajian tersebut ditata sedemikian rupa sehingga membentuk skema teoritis untuk menyelesaikan permasalahan dari masalah penelitian ini.

**BAB III : METODOLOGI PENELITIAN**

Pada bab ini menjelaskan tentang kerangka kerja penelitian, metode pengumpulan data, dan penjelasan terkait metodologi yang digunakan untuk menyelesaikan masalah dengan memanfaatkan analisis sentimen menggunakan metode *Naïve Bayes* dan *Information Gain*, serta alat bantu yang digunakan semasa mengerjakan penelitian ini.

**BAB IV : ANALISIS DAN HASIL**

Menyajikan analisis dan perancangan sistem metode klasifikasi *Naïve Bayes* dan seleksi fitur menggunakan metode *Informaton Gain*, perancangan pengujian dan manualisasi untuk menyajikan model hasil dari metode yang digunakan. Dan akan ditampilkan hasil dari analisis dan bentuk visualisasi analisis dari tools WEKA yang digunakan.

**BAB VI : PENUTUP**

Bab ini yang berisikan kesimpulan-kesimpulan yang diambil dari hasil analisis serta saran-saran yang mencakup keseluruhan dari hasil penelitian.