

BAB I

PENDAHULUAN

1.1 Latar Belakang

Teknologi *Big Data* menjadi fokus penelitian pada saat ini. Penggunaan *Big Data* pada umumnya digunakan untuk mengumpulkan beberapa data-data penting dalam satu teknologi yang nantinya bisa diolah. Salah satu jenis contoh data yang terkumpul dengan cepat adalah data teks (*email*, status media sosial), data waktu, data lokasi, data sensor, dan lain-lain. Data yang besar tersebut sering dikenal dengan istilah “*Big Data*”. K. D. Cahyo (2018). *Big data* menurut Gartner Inc., (2013) didefinisikan sebagai data yang memiliki ukuran (*volume*), kecepatan (*velocity*), dan/atau ragam (*variety*) yang ekstrim, yang menuntut pemrosesan informasi yang cepat dan inovatif untuk mendukung pengambilan keputusan dan otomatisasi proses. Terdapat tiga jenis *Big Data* yaitu *structured data*, *unstructured data*, *semi-structured data*.

Sedangkan menurut Torabi Asr & Taboada (2019) *Big data* merupakan media penyimpanan dalam bentuk besar.. Lalu, Dumbill (2012) menerangkan bahwa *Big Data* adalah data yang melebihi proses kapasitas dari konvensi sistem database yang ada. Data terlalu besar dan terlalu cepat atau tidak sesuai dengan struktur arsitektur database yang ada. Untuk mendapatkan nilai dari data, maka harus memilih jalan alternatif untuk memprosesnya.

Big Data dapat dianalisis untuk menjadi informasi atau pengetahuan yang berharga. Ada beberapa teknik dalam proses analisis *Big Data* antara lain, *Association Rule Learning*, *Classification tree analysis*, *Generic algorithms*, *Machine learning*, *Regression analysis*, *Sentimental Analysis*, dan *Social network analysis* (K. D. Cahyo, 2018).

Penelitian dalam lingkungan *Big Data* telah banyak dilakukan. Salah satu penelitian yang dilakukan oleh Olivandi et al (2018) yang melakukan pengujian analisis *Big Data* dengan Apache Spark pada *Big Data* berbasis HDFS, bahwa penggunaan Apache Spark sangat tepat karena dapat menurunkan *response time* rata-rata 50% sampai 70% dari Hadoop MapReduce. Namun, kelemahannya dalam optimalisasi kemampuan, membutuhkan spesifikasi *hardware* yang mumpuni.

Sementara itu penelitian yang dilakukan Demidova et al (2016) melakukan pengujian analisis *Big Data* dengan mengimplementasikan *Support Vector Machine (SVM)* dengan modifikasi *Particle Swarm Optimization*. Hasilnya menunjukkan Selama percobaan ditemukan bahwa *SVM ensembles* menunjukkan akurasi mulai dari 85,75% hingga 91,5%. Sementara keakuratan *SVM two level classifier* menunjukkan persentase sebesar 97,26%. Dengan demikian, *SVM two level classifier* meningkatkan akurasi klasifikasi hampir 3% dibandingkan untuk akurasi dari salah satu pengklasifikasi *SVM ensembles*.

Perbedaan hasil penelitian ini terjadi karena masing-masing peneliti memilih representasi data yang berbeda, sedangkan performa suatu algoritma sangat bergantung pada representasi dari data yang digunakan (Goodfellow et al., 2016).

Support Vector Machine (SVM) dikembangkan oleh Boser, Guyon, Vapnik, dan pertama kali dipresentasikan pada tahun 1992 di Annual Workshop on Computational Learning Theory. Konsep dasar SVM sebenarnya merupakan kombinasi harmonis dari teori-teori komputasi yang telah ada puluhan tahun sebelumnya, seperti margin hyperplane Duda & Hart (1973), Cover (1965), Vapnik (1964), dan sebagainya, kernel Aronszajn (1950) dan konsep-konsep pendukung yang lain. Belum pernah ada upaya merangkaikan komponen-komponen tersebut hingga tahun 1992 (L. B. D. Cahyo, 2018).

Penelitian yang memanfaatkan *Big Data* berbasis *cloud* masih belum populer di Indonesia khususnya dalam proses analisis *Big Data* pada *cloud-based data-warehousing* dengan algoritma *machine learning*. Disamping karena keterbatasan piranti lunak serta keterbatasan akses terhadap *Big Data* itu sendiri. Maka dari itu pada penelitian ini akan membahas tentang “Studi dan implementasi snowflake untuk analisis *big data* pada *cloud data warehousing* dengan menggunakan algoritma *support vector machine*”.

1.2 Rumusan Masalah

Berdasarkan uraian tersebut, maka rumusan masalah pada penelitian ini sebagai berikut:

1. Bagaimana menganalisis data berskala besar menggunakan *Snowflake* pada *cloud-based data-warehousing* dengan menerapkan algoritma SVM?
2. Bagaimana mengukur tingkat akurasi algoritma SVM dalam proses analisis *Big Data* berbasis *cloud-based data-warehousing*?

1.3 Batasan Masalah

Batasan masalah dalam penelitian ini adalah:

1. Analisis *Big Data* dibatasi pada *cloud-based data-warehousing* menggunakan Snowflake.
2. Penggunaan algoritma dibatasi hanya dengan menggunakan algoritma SVM pada Snowflake.
3. Pengukuran akurasi algoritma SVM pada *cloud-based data-warehousing* menggunakan Snowflake dengan variabel yang telah ditentukan.

1.4 Tujuan Penelitian

Tujuan penelitian tugas akhir ini adalah sebagai berikut:

1. Menganalisis data berskala besar dengan menggunakan Snowflake pada *cloud-based data-warehousing*.
2. Mengimplementasikan SVM pada Snowflake dalam proses analisis *Big Data*.
3. Mengukur tingkat keakuratan algoritma SVM dalam proses analisis *Big Data* berbasis *cloud-based data-warehousing*.

1.5 Manfaat Penelitian

Manfaat penelitian dari tugas akhir ini adalah sebagai berikut:

1. Menghasilkan sebuah model dari sebuah *Big Data* pada *cloud-based data-warehousing* yang nantinya data tersebut digunakan sebagai pengetahuan atau informasi.
2. Menghasilkan sebuah grafik dan persebaran dari pengimplementasi SVM pada proses analisis *Big Data*.

3. Menghasilkan tingkat akurasi dari algoritma SVM dalam proses analisis *Big Data* berbasis *cloud-based data-warehousing*.

1.6 Metodologi Penelitian

1. Studi Literatur

Studi pustaka dan literatur yang dilakukan, dipelajari, dan dikaji dalam menyelesaikan penelitian ini diambil dari buku, jurnal, internet, serta dokumen-dokumen yang berkaitan dengan penelitian.

2. Pengumpulan Data

Data yang dibutuhkan dalam penelitian ini berupa dataset yang telah diperoleh dari berbagai repositori yang berisi berbagai macam data untuk membantu penelitian yang akan dilakukan.

3. Perancangan Program

Perancangan program dibuat sesuai dengan analisis kebutuhan. Rancangan yang dibuat memuat proses memasukkan data dan inisialisasi data, proses pelatihan model, proses visualisasi dan analisis data, pengujian algoritma SVM pada Snowflake.

4. Implementasi Program

Implementasi dibuat untuk rancangan program yang telah disusun. Pada implementasi program, harus memuat proses memasukkan data dan inisialisasi data, proses pelatihan model, proses visualisasi dan analisis data,

pengujian algoritma SVM pada Snowflake seperti yang telah dijabarkan pada rancangan program.

5. Pengujian

Pengujian dilakukan untuk mengetahui kemampuan dari algoritma SVM serta model yang telah dibuat dengan beberapa parameter yang diuji. Hasil dari penelitian yang dilaksanakan berupa nilai akurasi dari SVM model.

1.7 Sistematika Penulisan

Sistematika penulisan yang digunakan dalam laporan tugas akhir ini adalah sebagai berikut:

BAB I PENDAHULUAN

Bab ini memaparkan mengenai gambaran umum yang meliputi latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, metodologi penelitian dan sistematika penulisan.

BAB II LANDASAN TEORI

Bab ini memuat uraian penelitian-penelitian terkait yang sudah pernah dilakukan dalam penelitian lain dan hubungannya dengan masalah penelitian yang sedang dilakukan. Selain itu, memuat dasar-dasar teoritis maupun penjelasan umum mengenai *Big Data*, algoritma *machine learning*, algoritma SVM, Snowflake, pendekatan analisis data, dan penjelasan umum lainnya yang berhubungan dengan penelitian.

BAB III METODOLOGI

Bab ini menguraikan mengenai alur penyelesaian masalah terhadap penelitian yang dilakukan. Selain itu juga menjelaskan analisis kebutuhan dan perancangan sistem yang akan digunakan. Pada bab ini dibahas mengenai alur penyelesaian masalah mulai dari studi literatur, diskusi dan konsultasi, pengumpulan data, perancangan program, implementasi program, sampai pengujian program.

BAB IV HASIL DAN PEMBAHASAN

Bab ini memuat uraian mengenai hasil dan proses pencapaian dalam menyelesaikan penelitian ini. Dimulai dari hasil studi literatur, diskusi dan konsultasi, pengumpulan data, perancangan program, implementasi program, sampai pengujian program.

BAB V KESIMPULAN DAN SARAN

Bab ini merupakan bab akhir dari penulisan laporan yang berisi mengenai simpulan-simpulan yang merupakan hasil analisis pada bagian sebelumnya serta saran yang perlu diperhatikan berdasarkan keterbatasan yang ditemukan.