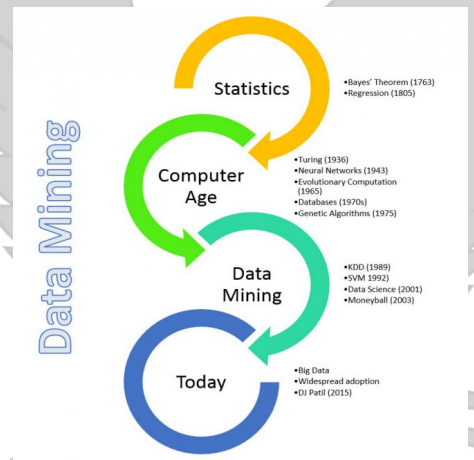
BAB I

PENDAHULUAN

1.1 Latar Belakang

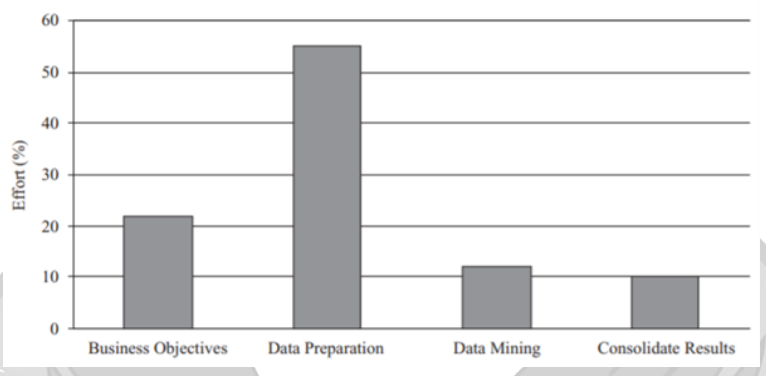
Perkembangan ilmu matematika dan statistik berperan penting dalam pertumbuhan proses pengolahan data. Pada Gambar 1.1, terlihat bahwa sejarah dari perkembangan pengolahan data sudah dimulai sejak tahun 1763 ketika Thomas Bayes mengeluarkan Teorema Bayes. Seiring dengan perkembangan ilmu matematika dan statistik, perkembangan teknologi juga ikut berkembang pesat. Perkembangan kedua hal tersebut membawa dampak positif dalam proses pengolahan data pada masa kini, kolaborasi antara ilmu yang ada dengan kecepatan memproses data menggunakan teknologi yang ada mengakibatkan lahirnya metode-metode baru dalam melakukan perhitungan.



Gambar 1.1: Sejarah perkembangan pengolahan data
Sumber: *Data mining techniques for herbs* [1]

Data mining adalah proses menganalisa kumpulan data dari berbagai perspektif dan menemukan korelasi dan pola untuk disimpulkan menjadi sebuah informasi yang berguna [2]. Pada Gambar 1.1, proses perkembangan data mining dimulai ketika Gregory Piatetsky-Shapiro mengeluarkan sebuah istilah bernama *Knowledge Discovery in Databases* (KDD). Sejak saat itu, banyak bermunculan metode dan algoritma baru dalam penganalisaan data yang sering disebut sebagai *machine learning*. *Machine learning* sendiri merupakan sebuah perpaduan antara teknologi dan ilmu matematika. Secara garis besar, *machine learning* terbagi menjadi dua bagian, yaitu *supervised learning* dan *unsupervised learning*.

Perbedaan pada kedua hal tersebut ada pada data yang digunakan, dimana pada *supervised learning* diketahui hasil dari data yang tersedia, sedangkan pada *unsupervised learning* tidak diketahui. Pada umumnya *supervised learning* digunakan untuk melakukan klasifikasi dan regresi, dan *unsupervised learning* digunakan untuk melakukan pengelompokan data.



Gambar 1.2: Persentase Upaya Dalam Proses *Data Mining*
Sumber: *Data Concepts, Models, Methods, and Algorithms* [3]

Perkembangan dari *data mining* memberikan peluang untuk melakukan perhitungan pada dataset dengan jumlah dan dimensi besar yang relatif sulit dan memerlukan waktu yang lama untuk dilakukan perhitungan secara manual. Langkah-langkah persiapan dalam proses *data mining* mencakup analisis dan spesifikasi jenis tugas *data mining*, dan pemilihan metodologi yang sesuai serta algoritma[3]. Meskipun banyak metode baru yang bermunculan, memilih metode yang tepat untuk sebelum melakukan perhitungan merupakan bagian yang relatif sulit dan memakan waktu, seperti yang dapat dilihat pada Tabel 1.2. Setiap metode memiliki keunggulan dan kekurangan tersendiri yang sangat bergantung pada data yang tersedia. Salah satu perbedaan pada data-data klasifikasi adalah jumlah kategori dari data hasilnya. Data klasifikasi yang hanya terdiri dari dua hasil adalah data *binary*, sedangkan data klasifikasi yang memiliki tiga atau lebih kategori disebut *multi-class*.

Dengan mengetahui permasalahan dan perbedaan pada jenis data klasifikasi yaitu *binary classification* dan *multi-class classification*, pada penelitian ini akan dicari metode yang lebih baik dalam melakukan klasifikasi pada jenis data yang berbeda, berdasarkan data yang tersedia. Random Forest dan Support Vector adalah metode yang dipilih untuk dibandingkan pada penelitian ini, karena kedua metode tersebut menggunakan pendekatan matematis yang cukup berbeda.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang ada, diharapkan penelitian ini dapat menjawab masalah-masalah berikut.

1. Bagaimana cara melakukan *binary classification*?
2. Bagaimana cara melakukan *multi-class classification*?
3. Bagaimana perbandingan performa *Random Forest* dan *Support Vector Machine* dalam melakukan klasifikasi?

1.3 Tujuan Penelitian

Tujuan penelitian adalah sebagai berikut.

1. Melihat proses perhitungan *Random Forest* dan *Support Vector Machine* dalam melakukan *binary classification*.
2. Melihat proses perhitungan *Random Forest* dan *Support Vector Machine* dalam melakukan *multi-class classification*.
3. Membandingkan Performa *Random Forest* dan *Support Vector Machine* menggunakan *Receiver Operating Characteristics*.

1.4 Batasan Masalah

Batasan masalah dan asumsi yang digunakan dalam mencapai tujuan penelitian ini adalah sebagai berikut.

1. Data diambil dari situs *archive.ics.uci.edu* dan diasumsikan valid.
2. Data yang diambil merupakan berasal dari bidang yang berbeda-beda (contoh: kesehatan, keuangan, benda)
3. Data yang tidak lengkap tidak akan digunakan pada penelitian ini.
4. Kernel yang digunakan pada *Support Vector Machine* adalah *Radial Basis Function*.
5. Pengolahan data menggunakan RStudio.
6. *Package* R yang digunakan untuk *Support Vector Machine* adalah e1071.
7. *Package* R yang digunakan untuk *Random Forest* adalah randomforest.

1.5 Manfaat Penelitian

Beberapa manfaat dari penelitian ini adalah sebagai berikut.

1.5.1 Manfaat Teoritis

1. Dapat diketahui proses pengklasifikasian dari *Random Forest* dan *Support Vector Machine*.
2. Dapat diketahui metode pengklasifikasian yang lebih baik di antara *Random Forest* dan *Support Vector Machine* pada jenis klasifikasi yang berbeda.
3. Sebagai bahan kajian untuk pengembangan dalam penelitian selanjutnya yang berkaitan dengan penyakit jantung, kanker, dan klasifikasi.

1.5.2 Manfaat Praktis

1. Mengetahui penggunaan praktis dari *Random Forest* dan *Support Vector Machine* dalam berbagai permasalahan yang membagi data menjadi beberapa kelompok yang ada di dunia nyata.
2. Membantu pembaca mengolah data klasifikasi menggunakan metode *Support Vector Machine* dan *Random Forest* bantuan software R studio.

1.6 Sistematika Penulisan

Struktur penulisan dari penelitian ini adalah sebagai berikut.

1. Pada Bab I dijelaskan mengenai latar belakang dilakukannya penelitian, rumusan masalah yang dihadapi, tujuan yang ingin dicapai melalui penelitian ini, batasan masalah yang digunakan dalam penelitian, manfaat dilakukannya penelitian, serta sistematika penulisan mengenai tugas akhir ini.
2. Pada Bab II dijelaskan teori dan definisi yang digunakan dalam penelitian sebagai referensi guna mencapai tujuan dari penelitian. Adapun teori yang akan diuraikan mencakup *Random Forest* dan *Support Vector Machine*.
3. Pada Bab III ditinjau langkah-langkah yang akan digunakan untuk setiap bagian dari proyeksi. Data-data yang akan digunakan juga dibahas dalam bab ini.

4. Pada Bab IV dilakukan simulasi, analisis hasil, dan pembahasan. Pada bab ini, akan diuraikan proses pengklasifikasian menggunakan metode *Random Forest* dan *Support Vector Machine* untuk mendapatkan model klasifikasi. Selanjutnya akan dilakukan perhitungan *error*.

5. Pada Bab V diberikan ringkasan dan konklusi dari hasil analisis yang diperoleh dan saran dalam melakukan analisis yang lebih efektif.