

## BAB I PENDAHULUAN

### 1.1 Latar Belakang

Pesatnya perkembangan teknologi dunia saat ini kini dapat membantu manusia dalam menyelesaikan berbagai masalah. Dengan ditemukannya penemuan-penemuan baru setiap tahunnya telah membuat perubahan pada pola hidup masyarakat. Salah satu penemuan yang dianggap penting yakni komputer, khususnya perangkat lunak atau software. Perangkat lunak yang ada kini dianggap telah banyak membantu pekerjaan yang dilakukan oleh manusia, baik dibidang kesehatan, olahraga, ekonomi dan yang lainnya. Perangkat lunak atau software ini juga dikembangkan secara berbeda-beda sesuai dengan tujuan pengaplikasiannya kelak. Fenomena ini juga dapat dirasakan dalam bidang pengklasifikasian. Klasifikasi dapat membedakan berbagai obyek baik makhluk hidup, pekerjaan, golongan serta yang lainnya.

Dalam dunia kesehatan, diagnosis penyakit pasien menjadi hal yang sangat sulit dilakukan. Namun demikian, catatan rekam medis telah menyimpan gejala-gejala penyakit pasien dan diagnosis penyakitnya. Hal seperti itu tentu sangat berguna bagi para dokter muda. Mereka bisa menggunakan catatan rekam medis yang sudah ada sebagai bantuan untuk mengambil keputusan tentang diagnosis penyakit pasien.

Kanker adalah pertumbuhan abnormal dari suatu sel atau jaringan dimana sel atau jaringan tersebut tumbuh dan berkembang tidak terkendali, kecepatan tumbuhnya berlebihan dan disertai perubahan perangai sel yang akhirnya mengganggu organ lain. Kanker terdiri dari beberapa jenis dengan nama-nama khusus yang diberikan tergantung pada sel-sel dan jaringan yang diserangnya. Salah satunya adalah kanker payudara, kanker ini sering terjadi pada wanita [TAG-02].

Jumlah perempuan penderita kanker payudara saat ini naik secara simultan, dan jumlahnya tersebut sangat menakutkan. Menurut data *National Cancer Institute* sejak tahun 1973 hingga 1998 angka penderita

kanker payudara telah naik lebih dari 40%. Pada tahun 2000, kurang lebih 182.800 perempuan didiagnosis menderita kanker payudara. Sebagian orang mungkin akan berpendapat bahwa kenaikan jumlah tersebut disebabkan oleh pendeteksian yang lebih awal dan lebih tepat. Namun, di kalangan perempuan berusia 80 tahun keatas pun dimana isu pendeteksian awal ini diragukan kebenarannya, angka penderita kanker payudara telah meningkat selama 30 tahun terakhir ini; yang awalnya 1 dari 30 perempuan sekarang menjadi 1 dari 8 perempuan. *The American Cancer Society* memperkirakan bahwa pada tahun 2000, 552.200 orang di Amerika Serikat akan meninggal akibat kanker, dan 40.800, atau sekitar 7 persen, di antaranya adalah perempuan penderita kanker payudara. Ini berarti 15% perempuan yang meninggal disebabkan oleh kanker payudara. Ini merupakan data statistik tahunan bagi rakyat Amerika Serikat. Namun lebih menakutkan lagi mengetahui bahwa sekitar 1.670.000 perempuan di dunia ini menderita kanker payudara [JOH-02].

Data kanker payudara hanya akan menjadi sekumpulan data yang tidak berguna jika tidak dilakukan penggalian data terhadapnya. Yang menjadi pertanyaan untuk data-data yang semakin menggenung jumlahnya dari waktu ke waktu itu adalah mau diapakan data-data tersebut. Apakah hanya untuk laporan akhir tahun, kemudian dibuang? Apakah hanya akan dikubur dalam gudang data dan dibiarkan? Tentu sayang sekali jika data-data tersebut tidak dimanfaatkan untuk keperluan perusahaan atau instansi itu sendiri. Banyak informasi terpendam yang dapat diambil dari sekumpulan data tersebut sehingga dapat memberikan suatu pengetahuan untuk penentuan kebijakan. Menggali data dapat dilakukan dengan cara mengklasifikasikan data penderita kanker payudara menjadi beberapa kelompok.

Dalam klasifikasi, terdapat berbagai jenis algoritma yang dapat digunakan dalam pengklasifikasian. Diantaranya adalah Bayesian classifier yang merupakan klasifikasi statistik yang dapat memprediksi keanggotaan class seperti probabilitas dari baris kedalam beberapa bagian class. KNN(K-Nearest Neighbour) yang merupakan sebuah metode untuk melakukan

klasifikasi terhadap obyek berdasarkan data pembelajaran yang jaraknya paling dekat dengan obyek tersebut. Dan juga algoritma SVM (*support vector machine*) yang merupakan algoritma komputer yang mempelajari dari contoh sebelumnya untuk menetapkan label untuk obyek. Pada penelitian yang dilakukan oleh J.Sreemathy (2012) dimana yang membandingkan antara algoritma naïve bayes ,KNN dan SVM. Hasil dari penelitian tersebut adalah algoritma Naïve Bayes dengan tingkat akurasi paling besar diantara algoritma lainnya sehingga dalam penelitian ini digunakan algoritma Naïve Bayes Classifier [SRE-12]

Naïve Bayes Classifier adalah salah satu dari teknik *data mining* yang paling populer dalam hal klasifikasi data set dalam jumlah besar. Naïve Bayes telah berhasil diaplikasikan kedalam masalah berbeda pada klasifikasi seperti peramalan cuaca, pengenalan pola serta gambar, diagnosis medis dan lainnya. Naïve Bayes Classifier telah banyak membantu dalam hal permasalahan *data mining*, namun Naïve Bayes masih dapat dikembangkan untuk meningkatkan performa serta akurasinya. Selain itu terdapat juga algoritma *clustering* lainnya, yaitu algoritma K-Means. K-Means adalah suatu metode penganalisaan data atau metode *data mining* yang melakukan proses pemodelan tanpa supervisi(*unsupervised*) dan merupakan salah satu metode yang melakukan pengelompokan data dengan system partisi. Metode K-Means berusaha mengelompokkan data yang ada kedalam beberapa kelompok, dimana data dalam satu kelompok mempunyai karakteristik yang sama satu sama lainnya dan mempunyai karakteristik yang berbeda dengan data yang ada di dalam kelompok yang lain. Dengan kata lain, metode ini berusaha untuk meminimalkan variasi antar data yang ada di dalam suatu klaster dan memaksimalkan variasi dengan data yang ada di klaster lainnya.

Sebuah penelitian yang dilakukan oleh Meiping (2009), yaitu penggabungan algoritma Naive Bayes dan K-Means untuk klasifikasi kartu kredit. Hasil dari penelitian tersebut menyimpulkan bahwa metode penggabungan algoritma Naive Bayes dan K-Means dapat digunakan untuk memperbaiki akurasi bayesian [MEI-09]. Sehingga pada penelitian ini akan menggunakan metode gabungan algoritma K-Means dan Naïve Bayes untuk

mengklasifikasikan kanker payudara menjadi 2 bagian yakni benigna dan maligna.

## 1.2 Rumusan Masalah

Berdasarkan latar belakang di atas maka ada beberapa rumusan masalah yang dapat diangkat, diantaranya:

1. Bagaimana menggunakan algoritma K-Means *Clustering* dan klasifikasi Naïve Bayes untuk data kanker payudara ?
2. Bagaimana akurasi yang dihasilkan dengan menggunakan algoritma K-Means dan Naïve Bayes ?
3. Bagaimana perbandingan akurasi yang dihasilkan K-Means dan Naïve Bayes dengan K-Means Konvensional?

## 1.3 Batasan Masalah

Sesuai dengan rumusan masalah yang telah diangkat maka dalam penelitian ini terdapat beberapa batasan masalah yang digunakan, yaitu:

1. Data yang digunakan adalah data set kanker payudara yang diambil dari UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer/breast-cancer-data>).

## 1.4 Tujuan

Berdasarkan rumusan masalah yang telah diangkat, adapun tujuan yang ingin dicapai dalam penelitian ini adalah :

1. Membangun sebuah aplikasi klasifikasi kanker payudara menggunakan algoritma K-Means dan Naïve Bayes.
2. Mendapatkan nilai akurasi aplikasi klasifikasi data kanker payudara yang menggunakan algoritma K-Means dan Naïve Bayes.
3. Membandingkan antara akurasi K-Means Naïve Bayes dan K-Means Konevensional.

## 1.5 Manfaat

Manfaat yang dapat diperoleh dari penelitian adalah dengan adanya aplikasi dari hasil implementasi algoritma K-Means dan Naïve Bayes ini dapat dijadikan sebagai salah satu solusi dalam permasalahan pendeteksian kanker payudara.

## 1.6 Sistematika Penulisan

Skripsi ini disusun berdasarkan sistematika penulisan sebagai berikut :

### 1. BAB I PENDAHULUAN

Bab ini berisi latar belakang, rumusan masalah, batasan masalah, tujuan dan manfaat penelitian, serta sistematika penulisan.

### 2. BAB II TINJAUAN PUSTAKA

Bab ini menguraikan teori-teori yang berhubungan dengan kanker payudara, data mining, klasifikasi, *clustering*, K-Means, Naïve Bayes dan gabungan K-Means dan Naïve Bayes.

### 3. BAB III METODOLOGI PENELITIAN

Bab ini membahas tentang data penelitian, perancangan sistem, perhitungan manual, perancangan antarmuka dan rancangan pengujian sistem.

### 4. BAB IV IMPLEMENTASI

Bab ini berisi spesifikasi perangkat lunak, implementasi sistem dan rancangan pengujian perangkat lunak yang dibangun.

### 5. BAB V PENGUJIAN DAN ANALISIS

Bab ini berisi pengujian serta analisis terhadap aplikasi yang dibangun.

### 6. BAB VI PENUTUP

Bab ini berisi kesimpulan dari seluruh rangkaian penelitian serta saran untuk kemungkinan pengembangan lebih lanjut.