

BAB I PENDAHULUAN

1.1 LATAR BELAKANG

Didunia nyata banyak sekali kasus ketiadaan nilai pada data untuk atribut tertentu atau hilangnya nilai yang terdapat pada beberapa aplikasi yang ada pada dataset. Permasalahan hilangnya nilai pada data yang bisa juga dikatakan data hilang. Unit maupun item terjadi pada sebagian survei skala besar karena kurang adanya respon yang menjadi penyebab data set hilang [1]. Data set yang hilang dapat terjadi pada dua tingkatan atau level yaitu unit atau item. Nonrespon pada tingkat unit terjadi ketika tidak ada informasi yang dikumpulkan dari responden. Contohnya jika responden menolak mengikuti survei atau tidak mengisi kuisioner. Nonrespon pada tingkat item merujuk pada ketidaklengkapan informasi yang diumpulkan oleh responden. Misalnya jika responden tidak mengisi satu atau dua pertanyaan dalam sebuah survei tetapi menjawab pertanyaan yang lain[2]

Data set yang hilang bukan menjadi masalah yang utama untuk studi yang bersifat substansif pada umumnya, tapi akan mengakibatkan masalah yang serius jika tidak ditangani secara tepat [3]. Pertama, data set yang hilang dapat memperlemah generalisasi dan menyebabkan potensi bias dalam estimasi parameter dari hasil penelitian. Kedua, mengakibatkan meningkatnya *standart error* dan kehilangan informasi yang berakibat berkurangnya *statistical power* jika diabaikan. Ketiga, data lengkap ditangani oleh semua prosedur statistik.

Selain itu juga banyak kasus yang ditemukan pada *database* yang setengahnya hilang dari data set. Sehingga sulit sekali untuk melakukan penambangan data, dimana hanya data yang lengkap untuk dilakukan analisis [4]. Data set yang hilang merupakan kelemahan umum yang banyak terjadi pada klasifikasi pola [5] dan merupakan masalah mempengaruhi hasil dari sistem prediksi data yang efektif [6]. Sehingga untuk memperkirakan nilai data yang hilang harus ada teknik penanganan khusus terhadap data hilang ialah sangat dibutuhkan.

Dengan demikian, dibutuhkan sistem yang lebih praktis dan mempunyai kemampuan dalam mengisi nilai yang hilang dengan nilai – nilai yang mungkin

berdasarkan informasi yang tersedia pada data set tersebut. Penelitian untuk mengatasi data set yang hilang telah banyak dilakukan oleh para peneliti dari waktu ke waktu. Dempster, Laird dan Rubin [7] menerapkan suatu pendekatan untuk diketahui berupa data tak lengkap untuk perhitungan secara iterasi dari estimasi *Maximum Likelihood* ketika observasi [8] memperkenalkan data yang hilang dengan memperkenalkan metode *bootstrap* [9].

Terdapat banyak metode untuk mengimulasi data set yang hilang. Metode *weighted k-Nearest Neighbor* merupakan pengembangan dari metode *k-Nearest Neighbor* (kNN) dan memiliki prinsip yang sama dengan metode *k-Nearest Neighbor* (kNN) yaitu mencari jarak terdekat antara data yang akan diuji dengan sejumlah k tetangga terdekatnya dalam data latih [10]

Beberapa penelitian menunjukkan bahwa klasifikasi *weighted k-Nearest Neighbor* (wkNN) adalah modifikasi dari klasifikasi *k-Nearest Neighbor* (kNN). Menurut Gou et al [11] permasalahan pada kNN terjadi ketika memilih k tetangga terdekat. Jika k yang dipilih sangat kecil, perkiraan klasifikasi cenderung menjadi tidak akurat karena data yang kurang, adanya noise, ambigu atau salah pelabelan. Nilai k yang terlalu besar dengan mudah membuat kinerja klasifikasi menurun karena adanya outlier dari kelas-kelas lain. Pemilihan k harus dioptimalkan untuk dapat meningkatkan akurasi. Menurut Pao et al. [12] pemberian bobot pada klasifikasi kNN dapat memperbaiki tingkat kesalahan. Pao et al. menyimpulkan hal ini dari hasil penelitian Dudani [13] mengenai classifier wkNN. Gou et al. menggunakan dataset *UCI Machine Learning Repository* <https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/>. Dataset yang digunakan merupakan dataset penyakit kanker payudara (*Breast Cancer Wisconsin*) yang diperoleh dari University of Wisconsin Hospitals, Madison dari Dr. William H. Wolberg. Pada data tersebut terdapat 699 data, 9 atribut, 16 data yang kosong dalam satu atribut, dan 2 kelas (jinak dan ganas) diperlukan perbaikan nilai supaya tidak terjadi sensitivitas pada data, juga digunakan pada penelitian Dudani sehingga dapat membandingkan kinerja kNN dan wkNN. Dataset UCI merupakan data seimbang. Pao et al. menggunakan data rekaman menggunakan Bahasa Mandarin yang juga merupakan data seimbang. Dari kedua penelitian tersebut,

dapat disimpulkan bahwa wkNN dapat mengatasi masalah sensitivitas kNN pada k tetangga terdekat.

1.2 PERUMUSAN MASALAH

Sesuai dengan uraian pada latar belakang diatas, maka yang menjadi masalah dalam tulisan ini adalah :

1.2.1 Permasalahan

Akibat dari hal tersebut adalah munculnya data set yang hilang yang menyebabkan data menjadi tidak lengkap. Data yang tidak lengkap akan menyebabkan potensi bias dalam estimasi parameter dan memperlemah generalisasi penelitian karena ukuran data yang berkurang sehingga mengurangi akurasi model.

1.2.2 Usulan Solusi

Metode yang digunakan untuk permasalahan yang ada dalam imputasi data set yang hilang menggunakan *weighted k-Nearest Neighbor* ialah untuk mencari jarak terdekat antara data yang diuji dengan sejumlah k tetangga terdekatnya dalam data latih.

1.2.3 Pertanyaan Penelitian

Bagaimana cara mengimputasi data hilang menggunakan metode *weighted k-Nearest Neighbor* ?

1.3 TUJUAN DAN MANFAAT

1.3.1 Tujuan Penelitian

Penerapan metode *weighted k-Nearest Neighbor* ini bertujuan untuk mengimputasi data set yang hilang dengan cara mencari jarak terdekat antara data yang diuji dengan sejumlah k tetangga terdekatnya dalam data latih.

1.3.2 Manfaat

Selain tujuan diatas, manfaat penelitian ini adalah nantinya dapat digunakan dengan ketika sebuah kasus data set yang hilang diabaikan maka dapat dipastikan akan menjadikan sulit memperoleh akurasi yang tinggi untuk hasil klasifikasi walaupun digunakan algoritma klasifikasi yang paling handal sekalipun.

1.4 BATASAN – BATASAN

Agar batasan dalam penelitian ini tidak terlalu luas, maka peneliti akan membatasi masalah-masalah sebagai berikut :

1. Sistem ini dikembangkan menggunakan metode *weihted k-Nearest Neighbor*.
2. Sistem bekerja untuk mengimpulasi data hilang dengan cara mencari jarak terdekat antara data yang diuji dengan sejumlah k tetangga terdekatnya dalam data latih.
3. Data berasal dari *UCI Machine Learning Repository* <https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/>. Dataset yang digunakan merupakan dataset *Breast Cancer Wiscosin* diperoleh dari University of Wisconsin Hospitals, Madison dari Dr. William H. Wolberg. Dengan 699 data, 9 atribut, 16 atribut nilai yang hilang dan 2 kelas (jinak dan ganas).

1.5 SISTEMATIKA PROPOSAL

Adapun sistematika pembahasan skripsi ini adalah sebagai berikut:

BAB I PENDAHULUAN

Bab ini berisikan latar belakang masalah, rumusan masalah, manfaat dan tujuan penelitian, batasan, masalah dan sistematika penulisan laporan.

BAB II STUDI PUSTAKA

Pada bab ini menjelaskan mengenai materi-materi yang menjadi landasan dalam penyusunan penelitian ini data set yang hilang, metode *k-Nearest Neighbor*, dan contoh perhitungan manual metode wkNN.

BAB III METODOLOGI USULAN

Bab ini menjelaskan mengenai flowchart dan rancangan dari sistem yang akan dibuat setelah melakukan analisa masalah, mulai dari kebutuhan software dan hardware, rancangan model serta alur dari sistem yang akan dikembangkan.