BAB I
PENDAHULUAN

1.1 Latar belakang

Data yang lengkap dan berkualitas sangat diharapkan ketika melakukan penelitian dan pengumpulan data.. Dalam suatu survei, pada umumnya tidak semua pertanyaan dijawab oleh responden. Ada berbagai alasan beberapa responden tidak menjawab beberapa pertanyaan sehingga terdapat nilai yang kosong atau informasi yang tidak tersedia sehingga membuat data penelitian tidak lengkap. Ketika data penelitian tidak lengkap maka hasil observasi tidak dapat dianalisis dengan baik. Pada beberapa aplikasi yang ada di dunia nyata terdapat banyak sekali kasus hilangnya nilai pada dataset atau ketiadaan nilai pada data untuk atribut tertentu. Permasalahan hilangnya nilai pada data ini lebih sering disebut dengan *missing data*. *Missing data* adalah suatu keadaan dimana terdapat nilai yang kosong atau nilai yang tidak lengkap dalam data. Pada suatu survei, missing data biasanya terjadi karena beberapa responden tidak menjawab satu atau lebih pertanyaan, hal ini disebabkan karena responden tidak ingin memberikan informasi yang dianggap pribadi dan rahasia. *Missing data* juga sering terjadi karena berbagai sebab, seperti kesalahan prosedur saat *entry data*, penyimpanan data yang kurang baik, dan berbagai sebab lain. Adanya *missing data* pada data yang akan diolah akan mengurangi tingkat keakuratan data tersebut saat akan diolah lebih jauh. *Missing data* merupakan kelemahan umum pada banyak skenario klasifikasi pola (García-Laencina et al., 2009) dan salah satu masalah yang dapat mempengaruhi hasil dari sistem prediksi data yang efektif (Malarvizhi, 2012). Sehingga penanganan terhadap *missing data* sangat penting dan dibutuhkan teknik penanganan khusus untuk memperkirakan nilai data yang hilang. Ada tiga mekanisme hilangnya data yaitu *Missing Completely at Random* (MCAR), Missing *at Random* (MAR), dan *Not Missing at Random*

(NMAR)(Tutz & Ramzan, 2015). MCAR adalah hilangnya nilai secara acak, dan MAR adalah terjadinya missing data juga secara acak, namun nilai yang hilang berhubungan dengan nilai pada *variabel* yang diketahui, tapi tidak pada *variabel missing data* itu sendiri. NMAR adalah *missing data* yang terjadi berhubungan dengan *variabel* yang mengandung *missing data* itu sendiri.

Banyak cara yang dilakukan untuk menangani kasus ini. Cara paling mudah adalah dengan menghapus baris data yang memiliki *missing data value* pada atributnya sebelum diolah. Pendekatan lain adalah dengan mengkonversi semua missing data value yang ada dalam data ke *value* yang baru yang nilainya sama dan dipilih secara *random* dari baris data yang lengkap. Ada juga yang memilih nilai yang paling sering muncul (*modus*) nilai tengah (*median*), atau nilai rata-rata (*mean*) dari data – data yang lengkap untuk mengisi *missing value*. Tapi tiga pendekatan ini hanya memiliki performansi yang bagus dalam menangani missing data value pada data – data yang berukuran kecil yang memiliki sedikit *missing data value* dan tidak berdampak besar pada pengolahan data selanjutnya.

Terdapat 3 metode yang dapat digunakan untuk penanganan missing data: *Case Deletion, Parameter Estimation*, dan *Imputation Techniques* (Rubin & Wiley, n.d.). Penilitian ini akan membahas *imputation techniques* (teknik imputasi) yakni metode penanganan *missing data* berdasarkan informasi yang tersedia pada *dataset* yang bertujuan untuk memprediksi nilai yang *valid* sebagai pengganti nilai yang hilang. *Missing data* akan menjadi masalah penting pada kasus klasifikasi dataset. Secara umum metode imputasi dan teknik klasifikasi adalah dua hal yang berbeda namun poin penting dari klasifikasi adalah bagaimana mendapatkan data pelatihan yang baik. Karena selain pemilihan metode yang tepat, akurasi hasil dari klasifikasi dipengaruhi oleh karakteristik dan kelengkapan *instance* dari sebuah data (Acu, 2004). Sehingga dengan melakukan imputasi terhadap *missing data* maka hal tersebut secara langsung dapat mempengaruhi hasil klasifikasi. Sebaliknya, ketika sebuah kasus missing data

diabaikan maka dapat dipastikan akan menjadikan sulit memperoleh akurasi yang tinggi untuk hasil klasifikasi walaupun digunakan algoritma klasifikasi yang paling handal sekalipun (Lai et al., 2019).

Beberapa penelitian menunjukkan bahwa penanganan missing data dengan menggunakan metode imputasi dapat meningkatkan akurasi klasifikasi dibandingkan dengan tanpa imputasi (Farhangfar et al., 2008) (García-Laencina et al., 2009) Karena nilai yang akan diimputasikan pada *missing data* didapatkan berdasarkan estimasi, sehingga dibutuhkan pemilihan metode imputasi yang tepat agar estimasi tersebut dapat mendekati data asli. Metode yang umum dilakukan ketika terdapat *missing data* adalah dengan membuang *missing data* tersebut. Namun, metode ini dapat menghilangkan informasi-informasi penting pada data yang kemungkinan ada pada data yang dihilangkan tersebut. Pengembangan dari metode imputasi telah banyak diteliti. Beberapa metode imputasi yang populer adalah : *Mean, Median*, klasterisasi, dan prediksi. Imputasi dengan menggunakan metode klasterisasi melakukan imputasi *missing data* dengan cara membagi dataset menjadi dua klaster, yaitu klaster yang berisi data dengan nilai komplit dan klaster yang berisi missing data. Selanjutnya, klaster yang berisi data komplit digunakan untuk mendapatkan nilai estimasi dengan cara menghitung nilai *Mean* atau *modus* seluruh data yang ada pada klaster tersebut. Nilai estimasi inilah yang akan digunakan untuk imputasi klaster yang berisi *missing data* (Malarvizhi, 2012)Sedangkan metode imputasi dengan model prediksi menggunakan sistem prediktif untuk memperkirakan nilai yang akan diimputasikan pada missing data (Malarvizhi, 2012).

Selain metode *mean*, *missing data* juga bisa diatasi dengan metode *K-Nearest Neighbour*. Dalam penelitian ini metode imputasi yang digunakan untuk mengatasi missing data adalah metode *K- Nearest Neighbour* (K-NN). K-NN melakukan imputasi *missing data* berdasarkan informasi dari observasi terdekat yang mempunyai kemiripan dengan *missing data.*

Penelitian ini akan memfokuskan diri pada implementasi salah satu metode imputasi yang dipakai untuk menangani missing value, yaitu metode *K – Nearest Neighbour Imputation* (KNN *Imputation*).

K - Nearest Neighbour Imputation (KNN Imputation) mengaplikasikan algoritma *K - Nearest Neighbour* (KNN) yang biasa digunakan pada proses klasifikasi untuk menangani *missing data*. Metode *K – Nearest Neighbour Imputation* (KNN Imputation) akan mencari nilai ketetanggaan terdekat dari suatu data yang memiliki missing data pada atributnya berdasarkan kalkulasi dari tetangga – tetangga terdekatnya sebanyak k yang diestimasi di awal.

1.2 Rumusan Masalah

Berdasarkan latar belakang, permasalahan yang dijadikan objek penelitian dan pengembangan penelitian ini adalah sebagai berikut :

- Bagaimana menangani variasi data yang memiliki *missing data value* pada data pelanggan indihome dengan menggunakan metode *K – Nearest Neighbour Imputation* (KNN Imputation) agar siap untuk diolah?
- Bagaimana menganalisis performansi dari dataset setelah menggunakan metode *K – Nearest Neighbour Imputation* (KNN Imputation)?

1.3 Batasan Masalah

Masalah yang akan dibahas memiliki batasan-batasan sebagai berikut:

- Data yang digunakan adalah data siap pakai dalam bentuk tabel yang berisi distribusi atribut dan *value*.
- Data yang digunakan tidak memiliki data error dalam distribusi *instance* nya.
- Data yang digunakan adalah data numerik.

1.4 Tujuan

Dalam penelitian ini , hal – hal yang ingin dicapai antara lain :

- Melengkapi missing data value yang ada dalam data yang bervariasi dengan menggunakan metode *K – Nearest Neighbor Imputation* (KNN Imputation).
- Menganalisis performansi dari dataset jika menggunakan metode *K – Nearest Neighbor Imputation* (KNN Imputation).

1.5 Manfaat Penelitian

- Memberikan metode alternatif yang dapat digunakan sebagai imputasi missing data, khususnya pada data Perusahaan Industri besar dan Sedang.
- Mengembangkan wawasan keilmuan mengenai metode imputasi dan penerapannya.