# 1. Pendahuluan

## 1.1 Latar Belakang

Perkembangan teknologi informasi saat ini berkembang sangat pesat seiring kebutuhan manusia yang bertambah pula. Salah satu kebutuhan manusia yang perkembangan cukup pesat adalah kebutuhan akan informasi. Dikarenakan hal tersebut maka semakin hari semakin banyak data yang menumpuk. Untuk mengimbangi perkembangan kebutuhan manusia akan informasi, maka perkembangan teknologi informasi khususnya di bidang manajemen data turut dikembangkan. Terkadang data yang terkumpul belum memberikan informasi secara jelas, maka diperlukan metode yang menangani pencarian informasi dalam data. Metode yang dapat digunakan untuk memebrikan solusi masalah tersebut adalah metode *data mining*.

Ping-Ning Tan dan Kumar Steinbach M. [12] menyatakan bahwa d*ata mining* adalah sekumpulan proses otomatis untuk menggali informasi yang berguna dari data yang besar, sehingga didapatkan informasi yang berguna.*Data mining* adalah bagian dari *Knowledge Discovery in Database* ( KDD ), yang keseluruhan prosesnya bertujuan merubah data mentah menjadi informasi yang bermanfaat[12]. Proses-proses dari KDD adalah dimulai dari *input data*, lalu *data preprocessing*, *data mining*, *postprocessing*, kemudian terakhir menghasilkan informasi. Dalam proses *preprocessing* bertujuan untuk merubah data mentah menjadi format data yang dibutuhkan untuk proses data mining[12]. Proses *preprocessing* memiliki beberapa cara yang dapat dilakukan, diantaranya adalah *data cleaning, data integration, data transformation,* dan *reduction*. *Data cleaning* adalah proses menghilangkan noise atau data yang tidak relevan untuk proses data mining[7]. *Data cleaning* sendiri memiliki beberapa cara yaitu pengisian *missing value*, identifikasi *outlier*, *smooting noisy data*, dan *correct inconsistent data.*

*Missing Value* adalah adalah data yang memiliki nilai atribut yang tidak lengkap atau hilang[12]. Dalam beberapa kasus, hal ini terjadi karena *human-error* atau kesengajaan subjek untuk tidak mengisi nilai tertentu, misal beberapa orang tidak ingin mengisi kolom umur atau berat badan[12]. Namun bagaimanapun juga, *missing value* harus diikutsertakan dalam proses analisis data. Beberapa cara menangani *missing value* dapat menggunakan beberapa pendekatan diantaranya adalah :(a) tidak menganggap objek yang mengandung *missing value;* (b) mengisi *missing value* secara manual; (c) ganti nilai pada *missing value* dengan nilai yang global atau rata-rata nilai dari objek; (d) isikan *missing value* dengan nilai yang memiliki kemungkinan terbaik[14]. Pendekatan (a) memiliki kelemahan kemungkinan kehilangan informasi lebih banyak, untuk pendekatan (b) memakan waktu dan biaya, sedangkan pendekatan (c) mengganti semua *missing value* dengan nilai yang sama sehingga dapat menyebabkan gangguan dalam distribusi data[14]. Pendekatan (d) adalah pendekatan yang memungkinkan meminimalis kekurangan dari pendekatan-pendekatan sebelumnya, namun tetap tidak menutup kemungkinan akan terjadi kesalahan dalam proses penggalian informasi nantinya. Penanganan *missing value* penting dilakukan untuk data yang membutuhkan hasil akurasi yang baik saat penggalian informasi. Hal ini disebabkan *missing value* dapat menjadikan hasil penggalian informasi pada data tersebut menjadi tidak valid. Contoh data yang membutuhkan tingkat akurasi yang tinggi adalah data kesehatan.

Teknik imputasi *missing value* dapat dibagi menjadi dua yaitu imputasi regresi berparameter (*parametric regression imputation*) dan imputasi non-parameter (*non-parametric regression imputation*)[14]. *Data parametric* adalah data yang dapat diregresikan secara linear atau fungsi umum lainya, sedangkan *data non-parametric* adalah data yang tidak dapat diregresikan secara umum [14]. CMI (Clustering-based Missing Value Imputation) adalah metode imputation regresi non-parameter (*non-parametric regression imputation*) berbasis clustering untuk menangani *missing value* yang memanfaatkan K-means dan fungsi kernel [14]. Pada CMI dilakukan beberapa proses dalam penanganan *missing value*, pertama lakukan strategi proses clustering yang dilakukan dengan metode clustering K-means. Kedua, dilakukan strategi imputasi fungsi Kernel atau disebut juga regresi non-parameter kernel imputasi. Fungsi Kernel ini juga merupakan metode yang efektif menangani *missing value* untuk komputasi yang efisien, kuat, dan stabil [14]. Berikutnya dengan algoritma CMI, *missing value* dapat diisikan berdasarkan cluster yang terbentuk. Singkatnya, setelah seluruh data ( termasuk missing value ) dibagi menurut *cluste*r yang sesuai, maka setiap *missing value* dari data diisikan dengan nilai yang mungkin berdasarkan cluster masing-masing data tersebut. Dengan menggunakan algoritma CMI akan dapat menangani *missing value* data yang bersifat non-linear atau yang tidak dapat difungsikan secara umum, sehingga bias didapatkan hasil akurasi yang lebih baik saat proses mining data.

## 1.2 Perumusan Masalah

Dari latar belakang permasalahan maka masalah dapat dirumuskan sebagai berikut :

1. Bagaimana akurasi CMI dalam penanganan *missing value* pada data numerik berdasarkan parameter RMSE?
2. Bagaimana pengaruh penanganan *missing value* menggunakan CMI terhadap proses data mining (klasifikasi) berdasar parameter *precision, recall,* dan *F-measure*?

Batasan masalah dari penelitian Tugas Akhir ini yaitu :

1. Jenis data adalah data numerik karena dibutuhkan perhitungan matematis dalam proses fungsi kernel.
2. *Clustering* dilakukan dengan metode *K-means clustering*.
3. Klasifikasi dilakukan dengan metode C-45 pada aplikasi Weka 3.6.6.
4. Untuk evaluasi dengan klasifikasi, data yang digunakan tidak ada *missing value* pada atribut kelas.
5. Dalam data set yang digunakan tidak ada *missing value* atau bernilai *null* karena *missing value* akan dibentuk secara acak.
6. Format file untuk pada proses penanganan *missing value* dengan CMI adalah .mat.
7. Format file untuk proses klasifikasi pada Weka adalah .arff.

## 1.3 Tujuan

Tujuan dari Tugas Akhir ini adalah sebagai berikut :

Mengimplementasikan dan mengetahui akurasi algoritma CMI dalam penanganan *missing value.*

## 1.4 Metodologi Penyelesaian Masalah

Metodologi yang digunakan dalam Tugas Akhir ini sebagai berikut :

1. Studi Literatur

   Pada tahap ini digunakan untuk mencari referensi dan data terkait data mining, missing value dan algoritma CMI. Referensi dapat berupa buku, website, paper, dan jurnal terkait yang akan dijadikan dasar teori.

2. Analisa kebutuhan dan perancangan perangkat lunak

   Tahapan untuk menentukan kebutuhan sistem, seperti identifikasi input, identifikasi output, identifikasi spesifikasi software dan hardware yang dibutuhkan.

   Dalam CMI, data input sebelumnya sudah melalui proses clustering menggunkan K-means, yang hasil dari clustering tersebut akan digunakan untuk lanjut ke proses berikutnya yaitu fungsi kernel. Setelah proses tersebut, maka dapat dilakukan missing value imputation. Setelah missing value ditangani, maka data dapat dilakukan proses mining sehingga diketahui akurasinya.

3. Implementasi perangkat Lunak

   Tahapan ini melakukan coding untuk pembangunan perangkat lunak penanganan *missing value* sekaligus analisis CMI algoritma yang diterapkan pada perangkat lunak.

4. Pengujian dan analisa perangkat lunak

   Tahap ini melakukan pengujian perangkat lunak dan analisis hasil pengujian. Pada tahap ini juga dilakukan evaluasi untuk mendapatkan hasil yang terbaik.

5. Pengambilan kesimpulan dan penyusunan laporan

   Tahap ini adalah tahap penarikan kesimpulan dari semua yang telah dilakukan dan penyelesaian penyusunan laporan. Penyusunan laporan dilakukan bertahap sejak awal pengerjaan Tugas Akhir.