-MOTTO-

"Janganlah kamu berduka cita, sesungguhnya Allah selalu bersama kita."

-QS At Taubah 40 –

"Wahai orang-orang yang beriman, jadikanlah sabar dan salat sebagai penolongmu. Sesungguhnya Allah beserta orang-orang yang sabar."

-QS Al Baqarah 153 -

"Maka sesungguhnya bersama kesulitan itu ada kemudahan."

- QS Al Insyirah 5-

"Barang siapa yang tidak mensyukuri yang sedikit, maka ia tidak akan mampu mensyukuri sesuatu yang banyak."

- HR. Ahmad –

"Adakalanya yang sedikit lebih berkah daripada yang banyak."

-Ali bin Abi Tholib-

"Manusia yang paling tinggi kedudukannya adalah mereka yang tidak melihat kedudukan dirinya, dan manusia yang paling banyak memiliki kelebihan adalah mereka yang tidak melihat kelebihan dirinya."

-Imam Syafi'i-

BAB I PENDAHULUAN

1.1 Latar Belakang

Data yang hilang adalah masalah umum yang terjadi di berbagai bidang seperti industri, sosial, ekonomi, politik, hiburan dan iklim (Farhangfar, 2008, Amanda, 2010, Bakhsh, 2014, dan Folguera, 2015). Permasalahan data sering kali muncul disebabkan beberapa hal diantaranya pengumpulan data pengamatan tidak berjalan dengan baik, pengukuran yang mungkin tidak lengkap, permasalahan responden menolak untuk menjawab beberapa pertanyaan tertentu dalam survey. Permasalahan yang mengakibatkan adanya nilai yang hilang disebut dengan *Missing value* (Wang, 2014). Nilai yang hilang dalam suatu set dianggap bermasalah jika data memiliki pengaruh besar pada pengambilan keputusan. Mekanisme terjadinya *Missing value*terbagi menjadi tiga yaitu: *Pertama*, *Missing Completely at Random* (MCAR) yang berarti bahwa terjadinya *Missing value*tidak berkaitan dengan nilai semua variabel, apakah itu variabel yang mengandung *Missing value*atau variabel yang terobservasi. *Kedua*, *Missing at Random* (MAR) yaitu terjadinya *Missing value*hanya berkaitan dengan variabel respon atau variabel pengamatan (Little dan Rubin, 1987). *Ketiga*, *Not Missing at Random* (NMAR) bahwa terjadinya *Missing value*pada suatu variabel berkaitan variabel respon dan variabel prediktor dataset (Pigott, 2011).

Proses *Machine Learning* (ML) pada data membutuhkan teknik *preprocessing* untuk menangani kekurangan karakteristik data yang berupa adanya nilai yang hilang dan ketidak konsistenan data. *Preprocessing* data yang hilang dapat diatasi dengan teknik imputasi yaitu proses estimasi pada data yang hilang, dengan nilai hasil estimasi digunakan untuk mengganti nilai data yang hilang tersebut (Little dan Ruben, 1987, 2002). Metode umum yang dipakai dalam ML untuk mengatasi nilai yang hilang diantaranya: *list wise deletion* yaitu menghapus variabel hasil observasi yang terdapat nilai yang hilang, sehingga mengakibatkan *error* yang lebih besar karena jumlah sampel berkurang. Teknik penghapusan data dianggap kurang tepat secara empiris, sehingga dikembangkan teknik imputasi dengan *Mean, modus, median* dan *hot deck* (Baraldi, 2010 dan Dong, 2013). *Mean Imputation* adalah nilai rata-rata fitur pada keseluruhan data untuk mengganti nilai yang hilang pada fitur yang sama. *Median Imputation* digunakan untuk menemukan nilai tengah dari semua nilai data hasil observasi untuk menggantikan nilai lainnya. *Modus Imputation* memanfaatkan nilai kemunculan data untuk mengganti

nilai yang hilang. *Modus Imputation* digunakan untuk memperbaiki *Mean Imputation* dan *Median Imputation* pada set data diskrit karena nilainya lebih beragam. *Hot deck* dikembangkan untuk memperbaiki metode imputasi *Mean*, *Median* dan *Modus* dengan memanfaatkan nilai *similarity* pada setiap nilai fitur untuk mengganti nilai yang hilang (Devi, 2015).

Permasalahan imputasi dapat diatasi dengan teknik *Mechine Learning* (ML), diantaranya teknik *clustering imputation, classification imputation dan computational imputation*. Teknik *clustering imputation* untuk mengatasi *Missing value*dengan memanfaatkan nilai bobot yang diperoleh dari pengelompokan data diantaranya *K-Means Imputation* (KMI), *Self Organizing Map Imputation* (SOMI), dan Fuzzy C-*Mean*s *Imputation* (FCMI). Metode *clustering imputation* tersebut menggunakan metode pembelajaran tanpa pengawasan untuk pengelompokan data berdasarkan kemiripannya. *K-Means Imputation* untuk estimasi nilai yang hilang berdasarkan nilai rata-rata bobot *centroid* pada pengelompokan data yang sama (Mehala 2008, dan Bakhsh, 2014). Metode KMI memiliki kelemahan yaitu melakukan inisialisai K secara random sehingga menyebabkan bobot yang di hasilkan estimasi berbeda-beda. Hasil nilai estimasi KMI yang diperoleh acak mengakibatkan penginisialisasi kurang baik, sehingga pengelompokan data lengkap menjadi tidak optimal (Jinhua, 2016). SOMI memiliki keunggulan cocok untuk data dengan atribut campuran (*multivariate*), karena menggunakan peluang alternatif bobot *centroid* pada *Best Machine Unit* (BMU) terkecil. Bobot tersebut menggunakan beberapa kriteria yang dihitung berdasarkan fungsi jarak. Hasil akhir cluster berdasarkan kemiripan anggota, diperoleh sampai tidak ada perubahan anggota objek data di klaster tersebut (Mehala, 2008, Yamaguchi, 2008, Peng, 2007, Kang, 2012 dan Vatanen, 2015, Saitoh, 2016, Khotimah, 2018).

Beberapa metode classification *imputation* yaitu dengan menggunakan *Decision Tree Imputation* (DTI), *Naïve Bayes Imputation* (NBI), *Support Vektor Machines Imputation* (SVMI), *k-Nearest Network Imputation* (kNNI), *Neural Network Imputation* (NNI), dan *Random Forest Imputation* (RFI). Metode klasifikasi untuk imputasi data yang hilang memiliki kelemahan tergantung dari pengolahan data awal (*preprocessing)* dan proses pencarian pola yang beragam sehingga waktu komputasi lebih lama ketika ketika proses pembelajaran (Pedro, 2010 dan Alireza, 2008). Metode DTI memiliki kekurangan yaitu kesulitan mendeteksi pola, ketika kelas-kelas dan kriteria yang

digunakan jumlahnya sangat banyak. Proses imputasi dengan DTI dapat menyebabkan nilai pengganti sama tetapi memiliki kelas yang berbeda. Selain itu, proses pengenalan pola *multidimensi* dan *multiclass* dapat meningkatkan jumlah memori pada saat pembelajaran (Farid, 2014).

Metode klasifikasi imputasi digunakan untuk mengukur hubungan antar data adalah *k-Nearest Network Imputation* (k-NNI) tergantung nilai *k* untuk mengukur tingkat kemiripan data. Semakin meningkat penggunaan nilai *k* akan menyebabkan hasil klasifikasi menjadi lebih kabur. Kelemahan metode k-NNI adalah hasil pembelajaran dipengaruhi jumlah dan bobot fitur yang sesuai, sehingga mengakibatkan komputasi yang lama dan bias yang rendah (Batista, 2003 dan Malarvizhi, 2012). SVMI dapat mengatasi masalah klasifikasi linier ataupun nonlinier, tetapi kesulitan menentukan nilai parameter yang optimal pada data dimensi besar (Pelckmans, 2005, Wen, 2014). *Neural Network Imputation* cocok digunakan pada data *multivariate*, tetapi membutuhkan preprosesing data terlebih dahulu dan kesulitan dalam menentukan parameter yang optimal. NBI dapat digunakan pada sejumlah kecil data pelatihan untuk memprediksi ouput dan cocok untuk data *multivariate* (Priya, 2005, Leng, 2009, Khotimah, 2018). Hasil kinerja NBI secara empiris dengan membandingkan proses imputasi dengan *Mean*, *modus*, *median* menghasilkan akurasi yang lebih tinggi untuk berbagai jenis data publik (Khotimah, 2019).

Optimasi untuk mengatasi masalah imputasi (computational *imputation*) diantaranya adalah *Genetic Algorithm Imputation* (GAI), *Ant Colony Optimization Imputation* (ACOI), dan *Particle Swarm Optimization Imputation* (PSOI) (García, 2011 dan Lobato, 2015). GAI dipilih untuk estimasi nilai *missing value*, karena metode ini dapat mengkodekan masalah dengan berbagai cara dari sekumpulan kadidat solusi (Shahzad, W., 2017). Algoritma komputasi lainnya, PSOI menggerakan partikel sebagai solusi untuk estimasi nilai yang hilang dengan menggunakan fungsi tertentu berdasarkan posisi dan kecepatan dari partikel. Algoritma PSOI memiliki kesulitan memprediksi ketepatan solusi awal pada sebagian besar data nonlinear, sehingga mengakibatkan terjadinya bias (Rezaie, 2006 dan Najib, 2017). *Missing value*pada data dimensi kecil sangat cocok menggunakan ACOI. Jika dimensi data semakin meningkat, ACOI menggunakan pertukaran informasi lebih lama, karena harus melewati semua jalur semut untuk mencapai solusi yang optimal (García, 2011).

Permasalahan umum pada Mechine Learning membutuhkan pengolahan data awal untuk meningkatkan klasifikasi (Liu, 2016). Pengolahan data awal pada data yang hilang ada dua perlakuan yaitu normalisasi dan teknik imputasi. Normalisasi pada data *multivariate* sangat cocok dengan menggunakan *z-score,* karena menghasilkan nilai rentang yang seragam (Box,1973). Teknik imputasi berbasis Artificial Intelligent (AI) memiliki kemampuan melakukan estimasi berdasarkan kesamaan pola, baik dengan cara pengelompokan maupun prediksi. Hasil estimasi diperoleh dengan melatih model pembelajaran hingga mencapai solusi maksimal berdasarkan karakteristik data. SOMI berbasis AI adalah teknik imputasi dengan mengestimasi bobot yang dihasilkan dari hasil akhir *learning* yang optimal pada data *multivariate* (Saitoh, 2016, Ahmad, 2016). Penelitian ini mengembangkan *Clustering* SOMMI (*Self Organizing Map Multiple Imputation*) dengan memanfaatkan vektor bobot akhir yang diperoleh dalam beberapa pembelajaran akhir. SOMMI pada umumnya menggunakan bobot berulang dan mencari nilai terbaik dari setiap bobot. Perbedaan SOMMI dengan SOM Single Imputataion (SOMSI) adalah terletak pada penggunaan bobot *centroid* terakhir dalam pembelajaran akhir dengan batas nilai kesalahan tertentu. SOMSI menghasilkan nilai imputasi yang kurang tepat karena hanya menggunakan satu bobot akhir, sehingga membutuhkan proses pencarian bobot berulang diantara bobot yang lain dari pembelajaran terbaik. Untuk mengatasi kelemahan SOMSI, maka melakukan perbaikan dengan SOMMI dengan menggunakan bobot rata-rata dari sejumlah bobot dalam iterasi akhir. Penggunaan SOMMI dengan berbagai bobot memiliki kelebihan untuk mengisi data secara berulang kali untuk menangani kompleksitas data yang sulit ditangani secara tepat (misalnya, data campuran) dengan pendekatan non-linear untuk mengatasi atribut kontinu dan kategori (Khotimah, 2018).

Naïve Bayes Classification (NBC) memiliki kelemahan yaitu sensitif terhadap nilai pengganti ketika nilai fitur yang hilang berupa atribut campuran. Permasalahan NBC pada data yang hilang terkadang muncul pada pola datanya, sehingga mengakibatkan *error* yang dihasilkan semakin meningkat. Ketika proses perbaikan NBC pada data yang hilang, membutuhkan pengisian data awal menggunakan metode umum yaitu *Mean*, modus, median hingga diperoleh data lengkap (Priya, 2017 dan Devi**,** 2015). Hasil estimasi nilai yang hilang dengan NBI menyebabkan terjadinya duplikat data. Hal ini disebabkan lebih dari satu nilai digunakan secara bersamaan pada fitur yang berbeda,

untuk menghasilkan data yang lengkap (Ludtke, 2016). Model hybrid *Self Organizing Map Imputation Naïve Bayes* (SOMINB) dikembangkan untuk memodelkan data campuran yang mengandung *Missing value*untuk memperbaiki kinerja NBC dengan memanfaatkan nilai bobot varian SOM sebagai pengganti nilai yang hilang (Khotimah, 2019). Nilai probabilitas NBC tergantung dari karaktersitik dan jumlah fitur. Semakin banyak jumlah data mengakibatkan kesulitan dalam memprediksi berdasarkan nilai probabilitasnya (Leng, 2009, dan Khotimah, 2019).

Penelitian ini akan mengembangkan *model Hybrid Naive Bayessian* Berbobot dengan Algoritma genetika pada pemodelan imputasi pada *Missing value*campuran yang sebut dengan *Self Organizing Map Imputation Genetic Algorithm Naïve Bayes* (SOMIGANB) (Khotimah, 2020). Algoritma genetika digunakan untuk perbaikan NB dengan pemilihan fitur yang menggunakan informasi heuristik pada data untuk mendapatkan set fitur yang optimal. Algoritma genetika dapat mengurangi jumlah atribut dalam data dimensi tinggi tanpa mengurangi informasi dari data. Algoritma genetika dikembangkan untuk menangani masalah dengan data dimensi tinggi untuk mengurangi waktu pemrosesan sehingga hasilnya lebih cepat diperoleh (Dharmistha, 2012). Pemilihan fitur dengan GA untuk meningkatkan kinerja klasifikasi dengan meningkatkan jumlah populasi melalui modifikasi ukuran kromosom dan penentuan parameter yang sesuai, diantaranya dengan menseleksi fitur yang digunakan pada klasifikasi (Meesad, 2008), NN, SVM (Pelckmans, 2005), *Decision Tree* (Farid, 2014), dan *Naïve Bayes* (Garcia, 2005, Patterkari, 2012, Kim, 2005, Peng, 2010, Tan, 2012, dan De Oliveira, 2018).

Model hybrid SOMIGANB diterapkan pada data medis Hepatitis Kronik yang diunduh dari repository UCI learning. Dataset ini berisi sejumlah atribut gejala medis beserta identifikasi untuk menentukan apakah penderita hepatitis hidup (live) atau mati (die) jika memiliki gejala-gejala berdasarkan fitur (atribut). Jumlah data sebanyak 155 *record*, dengan atribut yang menunjukkan gejala berjumlah 19 dan 1 atribut sebagai kelas. Atribut kelas berisi nilai 1 untuk “die” dan 2 untuk “live”. Proses penghilangan data pada atribut yang tidak tepat menyebabkan keputusan akhir terkadang salah. Data yang hilang berasal dari informasi yang tidak diperoleh karena adanya data pasien yang tidak lengkap seperti berkas data kesehatan tidak lengkap, pasien tidak mengisi jenis kelamin, tanggal lahir pasien, dan sebagainya (Anomaly, 2018). *Missing value*dalam dataset pasien berasal

dari data-data yang sebagian atributnya tidak memiliki nilai. Metode imputasi untuk preprosesing SOMI dan seleksi fitur pada data yang hilang digunakan untuk menghasilkan data lengkap dan sekaligus memilih fitur terbaik, sehingga dapat mengurangi dimensi data tanpa kekurangan informasi yang terkandung dalam kumpulan data tersebut.

Pembaruan bobot baru SOM memerlukan optimasi agar dapat meningkatkan efisiensi pembelajaran. Bobot akan mempengaruhi kualitas struktur utama dari jaringan yang kompleks dalam menemukan hubungan antara *node* dan *edge*. SOM mengalami ketidakkonsistenan seperti konstanta pembelajaran, algoritma pembelajaran, bobot awal. Penelitian ini melakukan pencarian bobot optimal SOM dengan menggunakan algoritma genetika. SOM berbasis NN bergantung pada algoritma *gradien* untuk mendapatkan bobot *model*. Selanjutnya, GA untuk seleksi bobot dengan memilih bobot terbaik pada setiap *cluster* data multidimensi, ketika *node output* tumbuh secara dinamis selama proses pengelompokan pola input. SOM dapat ditingkatkan dengan memilih unit yang menang dari node yang sesuai sehingga *cluster* stabil terbentuk. Dalam kasus data dimensi tinggi dengan nilai-nilai yang hilang, sejumlah besar metode pembelajaran mesin dapat diterapkan untuk menjelajahi area pencarian untuk imputasi dan pemilihan fitur dan parameter

1.2 Rumusan Masalah

Berdasarkan latar belakang diatas penelitian ini memiliki rumusan masalah sebagai berikut:

1. Bagaimana melakukan pengelompokan data dengan menggunakan SOMI untuk mendapatkan bobot yang tepat sebagai imputasi data yang mengandung *missing value*?
2. Bagaimana mengembangkan SOMI dengan melakukan SOMMI dengan mencari bobot pada learning yang paling sesuai untuk imputasi pada klasifikasi *Naïve Bayes*?
3. Bagaimana mengembangkan *model hybrid* SOMIGANB untuk seleksi fitur dan estimasi nilai *Missing value*pada klasifikasi NB data campuran?
4. Bagaimana menentukan bobot optimal SOMI untuk imputasi sebagai peprosesing data dengan menggunakan algoritma genetika?

1.3 Tujuan Penelitian

Penelitian ini memiliki tujuan untuk menjawab permasalahan diatas diantaranya:

1. Menentukan imputasi dengan bobot *Clustering* SOMI dengan mengelompokkan data yang mengandung *Missing value*untuk menghasilkan nilai bobot pengganti data yang hilang.
2. Mengembangkan SOMMI untuk menentukan bobot yang tepat dari beberapa learning SOMI yang menghasilkan bobot yang berbeda-beda untuk imputasi yang tepat yang dapat meningkatkan akurasi.
3. Mengembangkan hybrid klasifikasi *Missing value*data dengan algoritma genetika yang dapat diterapkan dalam data campuran. Sehingga *model* hybrid tersebut untuk mendapatkan fitur dan bobot fitur terbaik sebagai inputan dalam *Naïve Bayes.*
4. Menginisialisasi bobot SOMI dengan menggunakan algoritma genetika yang paling sesuai dengan fungsi *fitness* yang berbeda sesuai dengan variasi gabungan kondisi data.

1.4 Manfaat Penelitian

Penelitian ini diharapkan dapat bermanfaat dalam mengembangkan keilmuwan sebagai berikut:

1. Menambah keilmuan dari *Model Hybrid Naive Bayes* Berbobot dengan menggunakan pemetaan data dengan bobot vektor dan pemilihan atribut dengan menggunakan Algoritma genetika sebagai metode *alternative* dengan komputasi tinggi yang dapat digunakan untuk mengolah data yang ada *missing value*nya.
2. Mengembangkan *preprocessing* yang sesuai untuk data yang bermasalah baik data yang hilang maupun campuran untuk berbagai implementasi di bidang industri, kesehatan, pendidikan dan lainnya.