DATA CLEANSING PADA DATA RUMAH SAKIT

Ainayya Ghassani Lazuardy[1], Hari Setiaji S.Kom., M.Eng[2]
Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Islam Indonesia
Jalan Kaliurang Km 14, Ngemplak, Sleman, DI Yogyakarta, 55584
e-mail: [1]16523131@students.uii.ac.id, [2]hari.setiaji@uii.ac.id

ABSTRAK

Data dari Sistem Informasi Rumah Sakit tidak memiliki indikasi sebagai data yang berkualitas. Hal ini disebut juga dengan data kotor. Data kotor dapat berupa duplikasi data, tidak konsistennya data, dan data kosong. Data kotor ini dikarenakan kesalahan dalam entri data, skema sistem yang tidak berkesinambungan, data yang tumpang tindih dan lain – lain. Sehingga dibutuhkan tahapan Data Cleansing. Data Cleansing digunakan untuk mengubah data kotor menjadi data yang berkualitas yang nantinya data yang berkualitas akan diterapkan dengan tahapan data mining. Untuk itu, dibutuhkan metode yang tepat untuk melakukan proses Data Cleansing. Penelitian ini menerapkan metode metode Duplicate Elimination, Incosistency Detection, dan Handling Missing Entries pada data rumah sakit dan menghasilkan data rumah sakit yang sudah tidak mengalami duplikasi data, tidak konsistennya data dan tidak ada lagi data kosong.

Kata Kunci*: Data cleansing, Duplicate Elimination, Incosistency Detection, Handling Missing Entries*

1. PENDAHULUAN

Teknologi berkembang dengan pesat sehingga berdampak pada sektor kesehatan. Contoh teknologi dipakai dalam sektor kesehatan ialah Sistem Informasi Rumah Sakit. Sistem Informasi Rumah Sakit memiliki modul yang sesuai dengan standar pelayanan rumah sakit dan mudah dalam pengoperasian aplikasi. Sehingga pemanfaatan teknologi informasi ini merupakan solusi yang baik karena dapat meningkatkan kualitas pelayanan, efisisensi, dan penyediaan informasi secara cepat dan akurat [1].

Sistem Informasi Rumah Sakit dapat menyimpan data transaksi di rumah sakit seperti transaksi dalam pendaftaran pasien, pengobatan, sampai pengambilan obat. Data transaksi ini nantinya akan disimpan dalam suatu *database*. Sistem Informasi Rumah Sakit digunakan untuk mengakses data dengan lebih mudah daripada mencari data dalam tumpukan buku. Tidak butuh waktu lama dan praktis. Data yang disimpan tidak mudah rusak ataupun hilang daripada menggunakan catatan – catatan yang disimpan dalam kertas.

Namun data yang tersimpan di *database* rumah sakit masih kotor (dirty data). *Dirty data* ini diakibatkan oleh kesalahan sistem atau kesalahan pengguna (*human error*) dalam memasukkan data pada sistem atau perhitungan yang salah. Permasalahan pada *Dirty data* ditunjukkan dengan adanya kesalahan ejaan selama entri data, informasi yang hilang, duplikasi data atau data tidak valid lainnya [2]. Permasalahan ini penting untuk diselesaikan karena akan mempengaruhi kualitas suatu data karena menghasilkan informasi yang tidak akurat. Maka diperlukan pembersihan data terlebih dahulu atau yang dapat disebut dengan *data cleansing*. *Data cleansing* adalah suatu proses mendeteksi dan memperbaiki (atau menghapus) data set, tabel, dan database yang korup atau tidak akurat. Istilah ini mengacu pada identifikasi data yang tidak lengkap, tidak benar, tidak tepat, dan tidak relevan, yang kemudian dirty data tersebut akan diganti, dimodifikasi atau dihapus [3].

Penentuan metode yang cocok dalam mengatasi permasalahan pada *dirty data* sangat penting dan dibutuhkan dalam proses *data cleansing*. Penentuan metode *data cleansing* digunakan agar mendapatkan data yang berkualitas. Maka dari itu, penelitian ini bertujuan untuk menentukan metode *data cleansing* yang cocok untuk data rumah sakit agar data yang dihasilkan dapat diolah melalui proses *data mining.*

2. TINJAUAN PUSTAKA

Data Cleansing merubah data kotor menjadi data yang berkualitas agar dapat menghasilkan informasi yang akurat. Data yang berkualitas harus memiliki indikasi sebagai berikut :

1. Validitas : Tingkat kepatuhan terhadap aturan atau batasan bisnis yang ditetapkan
2. Akurasi : Tingkat kesesuaian ukuran dengan standar atau nilai sebenarnya
3. Kelengkapan : Sejauh mana semua tindakan yang diperlukan diketahui
4. Konsistensi : Sejauh mana seperangkat tindakan setara di seluruh sistem
5. Keseragaman : Tingkat pengukuran data yang ditetapkan dengan menggunakan satuan ukuran yang sama di semua sistem [4]

Permasalahan pada data seringkali terjadi dalam kumpulan data tunggal seperti *file* dan *database.* Permasalahan tersebut seperti kesalahan ejaan, informasi yang hilang atau data tidak valid lainnya. Biasanya disebabkan oleh petugas dalam entri data, model data dan desain skema sistem yang berbeda, atau data yang tumpang tindih, kontradiktif, dan tidak konsisten. Beberapa sumber data perlu diintegrasikan karena sumber data sering berisi data yang berlebihan dalam representasi yang berbeda.

Sehingga diperlukan untuk akses ke data yang akurat dan konsisten, konsolidasi berbagai representasi dan penghapusan data informasi duplikat [2].

Terdapat tiga metode data cleansing yang diterapkan, yaitu :

1. *Duplicate Elimination*

   Metode ini menentukan apakah dua data atau lebih merupakan representasi rangkap dari entitas yang sama. Cara kerja metode ini ialah dengan membandingkan setiap data yang ada berdasarkan atribut yang ditentukan. Maka hasil didapatkan adalah berapa banyak data yang mengalami duplikasi data. Lalu metode ini dapat menghapus duplikasi data sehingga hanya satu dari semua data duplikat yang disimpan [4].
2. *Incosistency detection*

   Penerapan metode ini dilakukan ketika data tersedia dari beberapa sumber yang berbeda dan dalam format yang berbeda. Tidak konsistennya data dapat disebabkan oleh kesalahan manusia (*human error*) dalam pengentrian data [5]. Misalnya pada atribut NO_POLIS seharusnya diisikan dengan nomor 6 digit. Namun karena adanya kesalahan dalam pengentrian data, ada data yang disimpan dengan 4 digit atau diisi dengan "-". Metode ini akan menemukan anomali yang terjadi pada data dan mengubahnya menyesuaikan format yang telah ditentukan.
3. *Handling Missing Entries*

   Dalam mengatasi data yang hilang (*missing entries*), metode ini dapat menghilangkan data apapun yang berisi entri yang hilang [5]. Semisal ada kasus pasien yang rawat inap pada suatu rumah sakit. Namun didata tersebut menunjukkan bahwa ID_BED tidak dientri sehingga data pasien tersebut dapat dihapus.

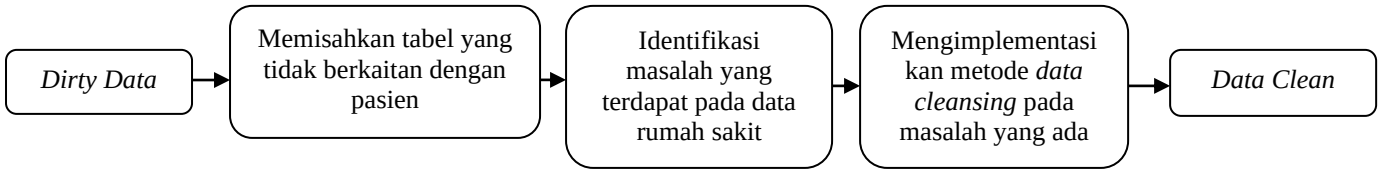
   Penelitian sebelumnya telah dilakukan untuk menghilangkan data kosong menggunakan metode Decision Tree karena metode ini adalah metode yang sangat dipahami oleh manusia [6]. Lalu penelitian yang dilakukan oleh [7] berhasil menghilangkan duplikasi data menggunakan metode the smith waterman algorithm dan the union find data. Selanjutnya [8] menggunakan metode Duplicate Elimination sebelum melanjutkan tahapan machine learning pada jaringan.

3. METODE PENELITIAN DAN KAKAS ANALISIS

3.1 Sumber Data

Data yang akan diolah dengan *data cleansing* berasal dari *database* sistem informasi rumah sakit di suatu rumah sakit di Indonesia. *Database* merupakan *database Oracle* dengan besar data 500 Mb. Terdapat 161 tabel dalam *database* tersebut. Batasan masalah pada penelitian ini ialah tabel yang digunakan untuk proses *data cleansing* hanya yang berkaitan dengan pasien.

3.2 Tahapan Data Cleansing



Gambar 1. Tahapan Data Cleansing

Pada tahap ini akan dilakukan analisis dan perancangan penelitian sebagai berikut :

1. *Dirty Data*

Data diambil dari suatu sistem informasi rumah sakit dengan *import* interlokal *database* secara manual. Lalu dalam tahapan data cleansing dilakukan *export* tabel dan ubah *file* menjadi excel.

2. *Memisahkan tabel yang tidak berkaitan dengan pasien.*

Sesuai dengan batasan masalah pada penelitian ini adalah mengambil tabel data yang hanya berkaitan dengan pasien. Maka dari 161 tabel hanya 14 tabel saja yang berkaitan dengan pasien. Tabel – tabel tersebut ialah :

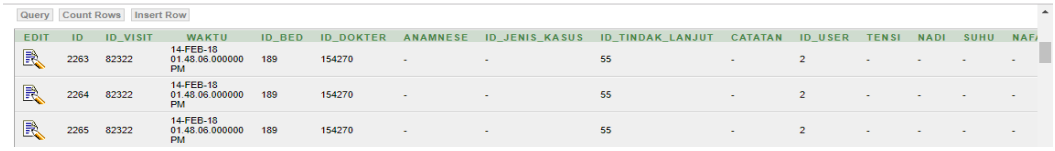
- ASURANSI_KEPESERTAAN_VISIT
- ANTRIAN
- RAWAT_JALAN
- BILLING
- RAWAT_DARURAT
- KUNJUNGAN
- KUNJUNGAN_BPJS
- LOG_BATAL_KUNJUNGAN
- PASIEN
- HASIL_PEMERIKSAAN_LAB
- PEMBAYARAN
- PENDUDUK
- PENJUALAN_RESEP
- VISIT

3. *Identifikasi masalah yang terdapat pada data rumah sakit*

Masalah yang terdapat pada data sistem rumah sakit ini ada tiga, yaitu

a. Duplikasi data

Duplikasi data ditemukan pada tabel RAWAT_JALAN, PENDUDUK, RAWAT_DARURAT. Penemuan duplikasi data ini dengan cara membandingkan tiap baris berdasarkan atribut satu dan yang lainnya. Sebagai contoh pada tabel RAWAT_JALAN. Ditemukan pada ID 2263 – 2265 mengalami duplikasi data karena isi data selain pada kolom ID sama. Hal tersebut dapat dikatakan sebagai duplikasi data karena data pada kolom WAKTU sama persis, sehingga tidak relevan.



Query | Count Rows | Insert Row

EDIT	ID	ID_VISIT	WAKTU	ID_BED	ID_DOKTER	ANAMNESE	ID_JENIS_KASUS	ID_TINDAK_LANJUT	CATATAN	ID_USER	TENSI	NADI	SUHU	NAFA
	2263	82322	14-FEB-18 01.48.06.000000 PM	189	154270	-	-	55	-	2	-	-	-	-
	2264	82322	14-FEB-18 01.48.06.000000 PM	189	154270	-	-	55	-	2	-	-	-	-
	2265	82322	14-FEB-18 01.48.06.000000 PM	189	154270	-	-	55	-	2	-	-	-	-

Gambar 2. Duplikasi data pada tabel RAWAT_JALAN

b. Tidak konsistennya data

Tidak konsistennya data ditemukan pada tabel ASURANSI_KEPESERTAAN_VISIT. Ditemukan bahwa pada kolom NO_POLIS ada yang berisikan dengan "-" selain angka dengan menelusuri relasi tabel. Hal ini tentunya akan merusak sistem karena NO_POLIS merupakan *foreign key.*

NO_POLIS
1018624228
191418554
-
-

Gambar 3. Tidak konsistennya data pada tabel ASURANSI_KEPESERTAAN_VISIT

c. Data kosong.

Sebagai sampel, data kosong ditemukan pada tabel RAWAT_JALAN dengan cara memilih kolom ID_BED dan menyamakan dengan data kosong. Pada kolom ID_BED dan ID_DOKTER terdapat data kosong dengan ditandai dengan "?". Seharusnya dalam tabel RAWAT_JALAN, kolom ID_BED dan ID_DOKTER berisi kan nomor ID dan kolom – kolom ini menjadi *foreign key*.

ID	ID_VISIT	WAKTU	ID_BED	ID_DOKTER
2233	82352	31-JAN-18 03...	186	52399
2172	82302	18-JAN-18 12...	?	?
2173	81934	12-DEC-18 1...	?	?
2178	13426	19-JAN-18 12...	?	?
2179	82302	19-JAN-18 12...	?	?
2180	13426	19-JAN-18 12...	?	?
2181	13426	19-JAN-18 12...	?	?
2182	13426	19-JAN-18 12...	?	?
2183	13426	19-JAN-18 12...	?	?
2191	82107	23-JAN-18 12...	374	154270

Gambar 4. Data kosong pada tabel RAWAT_JALAN

4. Mengimplementasikan metode data cleansing pada masalah yang ada

Dengan permasalahan yang ada pada data sistem rumah sakit dapat diatasi dengan menggunakan beberapa metode yang sudah disebutkan diatas. Pengimplementasian metode *data cleansing* menggunakan *RapidMiner* untuk pemrosesan data.

5. Data Clean

Setelah menggunakan metode *data cleansing* pada data maka akan menghasilkan data yang bersih dan berkualitas.

3.3 Kakas Analisis

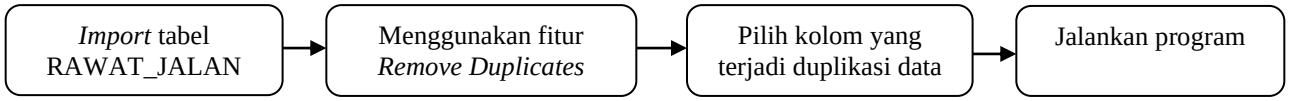
Penelitian ini menggunakan perangkat lunak RapidMiner dalam melakukan *data cleansing* pada data rumah sakit. Perangkat lunak RapidMiner adalah sebuah perangkat lunak yang memiliki kecerdasan buatan melalui platform sains data dan sebagai mesin *data mining*. RapidMiner dapat memberikan solusi dari persiapan data hingga pembelajaran mesin dan penerapan model prediksi [9]. Dalam penggunaan data mining, RapidMiner menyediakan prosedur *data mining* dan *machine learning* termasuk : ETL (*Extraction*, *Transformation*, *Loading*), *data preprocessing*, visualisasi, *modelling* dan evaluasi [10].

RapidMiner memiliki keunggulan diantaranya :

1. Merupakan *freeware*. Sehingga tidak perlu mengeluarkan biaya lebih untuk instalasinya.
2. RapidMiner menggunakan bahasa pemrograman java sehingga dapat digunakan di berbagai sistem operasi.
3. Memungkinkan untuk eksperimen data dalam skala besar.
4. Dapat menampilkan grafis yang canggih seperti diagram batang, 3D Scatter plots, dan lain – lain.
5. Memiliki beberapa *plug in* sehingga dapat menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik [10].

Penggunaan RapidMiner dalam mengatasi permasalahan yang terdapat pada data sistem rumah sakit :

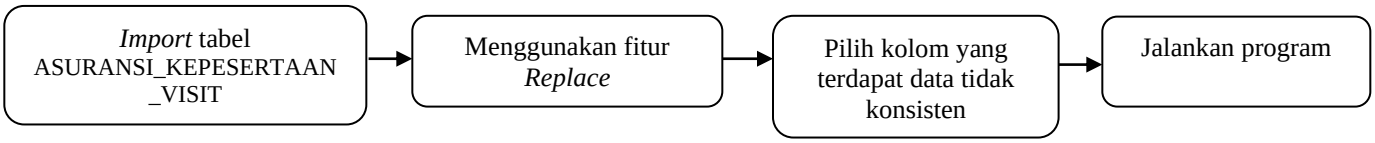
1. Menghilangkan duplikasi data.



Gambar 5. Tahapan menghilangkan duplikasi data pada RapidMiner

Dalam menghilangkan duplikasi data pada data rumah sakit, dapat menggunakan RapidMiner sebagai kakas analisis. Duplikasi data dapat dihilangkan dengan fitur *Remove Duplicates*. Sebagai contoh duplikasi data terjadi pada tabel RAWAT_JALAN. Maka sebagai langkah awal yang dilakukan adalah *import* file *.xls* menggunakan fitur *Read Excel* pada tabel RAWAT_JALAN. Lalu menggunakan fitur *Remove Duplicates*. Selanjutnya memilih nama kolom yang terjadi duplikasi data pada tabel RAWAT_JALAN, yaitu ID_BED, ID_DOKTER, ID_TINDAK_LANJUT, ID_VISIT dan WAKTU. Langkah terakhir, jalankan program.

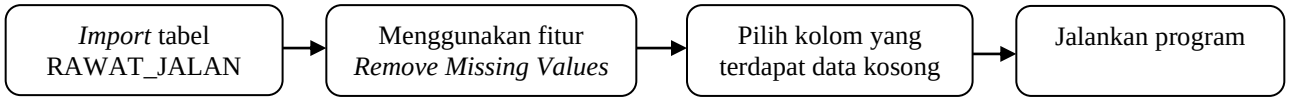
2. Membuat data menjadi konsisten



Gambar 6. Tahapan menghilangkan data tidak konsisten pada RapidMiner

Permasalahan data yang tidak konsisten terjadi pada tabel ASURANSI_KEPESERTAAN_VISIT pada kolom NO_POLIS. Dengan menggunakan RapidMiner, langkah pertama adalah *import file* .xls menggunakan fitur *Read Excel* pada tabel ASURANSI_KEPESERTAAN_VISIT. Selanjutnya data tidak konsisten dapat diatasi menggunakan fitur *Replace*. Lalu ubah data "-" menjadi NO_POLIS yang *default* yaitu, "123456789". Langkah terakhir, jalankan program.

3. Menghilangkan data yang kosong.



Gambar 7. Tahapan menghilangkan duplikasi data pada RapidMiner

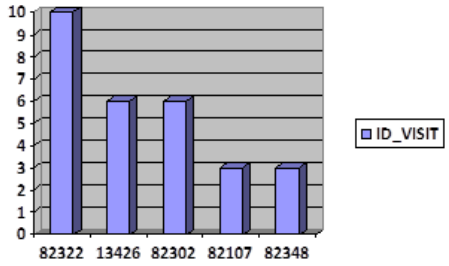
Sebagai contoh data kosong terjadi pada tabel RAWAT_JALAN. Data kosong dapat dihilangkan dengan menggunakan fitur *Remove Missing Values* pada RapidMiner. Langkah pertama adalah *import file .xls* menggunakan fitur *Read Excel* pada tabel RAWAT_JALAN. Lalu menggunakan fitur *Remove Missing Values* dan masukkan atribut yang terdapat data kosong seperti ID_BED dan ID_DOKTER. Langkah terakhir, jalankan program.

4. HASIL DAN PEMBAHASAN

Hasil yang didapatkan adalah sebagai berikut :

1. Menggunakan metode *Duplicate Elimination* untuk mengatasi duplikasi data.

Data pada tabel RAWAT_JALAN menunjukkan mengalami duplikasi data pada ID_VISIT dan juga terjadi pada waktu yang sama pada kolom WAKTU. ID_VISIT dengan nomor "82322" paling banyak mengalami duplikasi data sejumlah 10 baris.



Gambar 8. Jumlah duplikasi data pada ID_VISIT lima terbesar

Tabel RAWAT_JALAN setelah menggunakan metode *Duplicate Elimination*.

ID	ID_VISIT	WAKTU	ID_BED	ID_DOKTER	ANAMNESE	ID_JENIS_K...	ID_TINDAK_...	CATATAN	ID_USER	TENSI	N
2257	82322	14-FEB-18 1...	189	154270	?	?	53	-	2	-	-

Gambar 9. Hasil menghilangkan duplikasi data

2. Menggunakan metode *incosistency detection* untuk mengatasi data yang tidak konsisten.

Pada tabel ASURANSI_KEPESERTAAN_VISIT terdapat NO_POLIS berisi "-" sedangkan yang lain berupa angka berjumlah 55 baris.

Row No.	EDIT	ID	NO_POLIS	ID_VISIT	ID_ASURAN...	ID_PERUSA...
1	?	113469	-	72736	13	?
2	?	113477	-	72729	11	?
3	?	113478	-	72729	15	?

ExampleSet (55 examples, 0 special attributes, 6 regular attributes)

Gambar 10. Data sebelum menggunakan metode Inconsistency Detection.

Setelah menggunakan metode *inconsistency detection,* data dengan "-" diubah menjadi "123456789".

Row No.	EDIT	ID	NO_POLIS	ID_VISIT	ID_ASURAN...	ID_PERUSA...
1	?	113469	123456789	72736	13	?
2	?	113477	123456789	72729	11	?
3	?	113478	123456789	72729	15	?

Gambar 11. Hasil dari metode Inconsistency Detection.

3. Menggunakan metode *Handling Missing Entries* untuk mengatasi data kosong.

   Data kosong ditemukan pada tabel RAWAT_JALAN. Terdapat data kosong pada atribut ID_BED berjumlah 8 baris. Dengan demikian, data tersebut dihilangkan.

Row No.	EDIT	ID	ID_VISIT	WAKTU	ID_BED	ID_DOKTER	ANAMNESE	ID_JENIS_K...	ID_TINDAK_...	CATATAN	ID_USER	TEN
1	?	2172	82302	18-JAN-18 12...	?	?	?	?	?	-	2	-
2	?	2173	81934	12-DEC-18 1...	?	?	?	?	53	-	2	-
3	?	2178	13426	19-JAN-18 12...	?	?	?	?	53	-	2	-

Gambar 12. Data kosong pada kolom ID_BED

5. KESIMPULAN

Data Cleansing pada data rumah sakit membutuhkan beberapa metode untuk mengubah data kotor menjadi data yang berkualitas. Ditemukan bahwa data sistem rumah sakit mengalami duplikasi data pada tabel RAWAT_JALAN berjumlah 10 baris, tidak konsistennya data pada tabel ASURANSI_KEPESERTAAN_VISIT berjumlah 55 baris, dan data kosong pada tabel RAWAT_JALAN dengan atribut ID_BED berjumlah 8 baris. Duplikasi data tidak hanya ditemukan pada tabel RAWAT_JALAN, tetapi ditemukan juga pada tabel PENDUDUK, RAWAT_DARURAT dan untuk tidak konsistennya data ditemukan juga pada tabel PENDUDUK, ANTRIAN, dan KUNJUNGAN. Selain itu, untuk tabel yang terdapat data kosong terdapat pada tabel RAWAT_JALAN, KUNJUNGAN, dan LOG_BATAL_KUNJUNGAN. Maka data rumah sakit ini menerapkan metode *Duplicate Elimination*, *Incosistency Detection*, dan *Handling Missing Entries*. Ketiga metode ini dapat dijalankan secara paralel atau berurutan. Dapat dijalankan untuk metode *Duplicate Elimination* terlebih dahulu lalu *Inconsistency Detection* dan setelahnya menggunakan *Missing Entries*. Sebaiknya dalam kasus pada data sistem rumah sakit ini lebih baik menggunakan metode *Duplicate Elimination* dan *Handling Missing Entries* dikarenakan banyaknya data yang bermasalah dalam duplikasi data dan data kosong. Menggunakan metode Duplicate Elimination didukung oleh penelitian dari [11] bahwa masalah mendeteksi dan menghilangkan data duplikat adalah salah satu masalah utama dalam pembersihan data dan kualitas data pada *database*. Penghapusan duplikat sulit karena disebabkan oleh beberapa jenis kesalahan seperti kesalahan tipografi, dan representasi berbeda dari nilai logis yang sama. Maka dari itu, dengan metode – metode tersebut data rumah sakit sudah tidak mengalami duplikasi data, tidak konsistennya data dan tidak ada lagi data kosong.

DAFTAR PUSTAKA

[1] A. Bustomi, "ARSITEKTUR INFORMASI SISTEM INFORMASI MANAJEMEN RUMAH SAKIT (SIMRS)," 2016.

[2] E. Rahm and H. H. DO, "Data Cleaning: Problems and Current Approaches," *IEEE Trans. Cloud Comput.*, vol. 2, no. 1, pp. 1–1, 2014.

[3] A. Riezka, I. Atastina, and K. Maulana, "Latar belakang Pendahuluan Data-Cleaning adalah suatu proses mendeteksi dan memperbaiki (atau Rumusan masalah Berdasarkan latar belakang diatas , permasalahan yang menjadi fokus pada," 2011.

[4] H. Müller and J. Freytag, "Problems , Methods , and Challenges in Comprehensive Data Cleansing," pp. 1–23, 2003.

[5] C. C. Aggarwal, *Data Mining*. New York: Springer, 2015.

[6] G. A. Liebchen, "Data Cleaning Techniques for Software Engineering Data Sets," *PhD Thesis*, no. October, 2010.

[7] A. Monge, "An adaptive and efficient algorithm for detecting approximately duplicate database records," *On-line Doc. URL http//citeseer. nj. nec. com/ …*, no. August, pp. 0–17, 2000.

[8] C. Ambedkar and V. K. Babu, "Detection of Probe Attacks Using Machine Learning Techniques," vol. 2, no. 3, pp. 25–29, 2015.

[9] RapidMiner, "About RapidMiner," 2019. [Online]. Available: https://rapidminer.com/us/.

[10] I. W. S. Wicaksana, *Belajar Data Mining dengan Rapid Miner*. Jakarta, 2013.

[11] J. J. Tamilselvi and C. B. Gifta, "Handling Duplicate Data in Data Warehouse for Data Mining," vol. 15, no. 4, pp. 7–15, 2011.